THE JOHNS HOPKINS UNIVERSITY
STUDIES IN EDUCATION

No. 3

EDITED BY

EDWARD F. BUCHNER

in cooperation with

C. MACFIE CAMPBELL

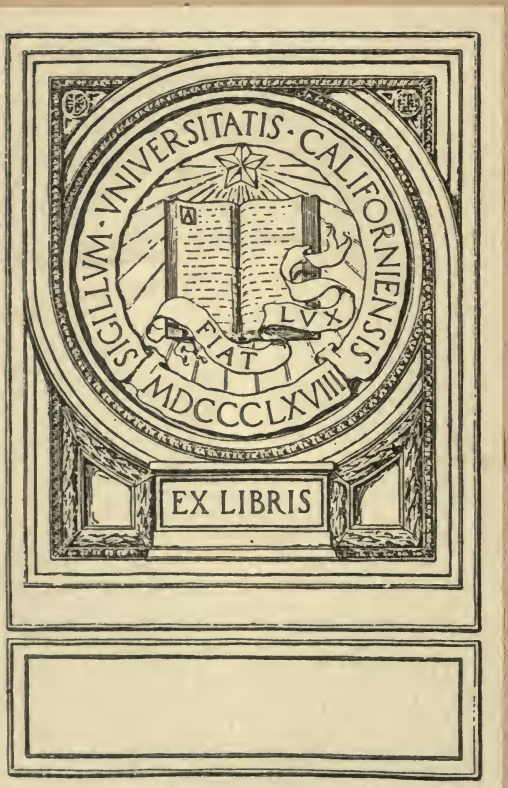
STUDIES IN EXPERIMENTAL EDUCATION

BY

BIRD T. BALDWIN, PH.D.

AND OTHERS

BALTIMORE
THE JOHNS HOPKINS PRESS
1920


SIGILLVM · VNIVERSITATIS · CALIFORNIENSIS
LVX
FIAT
MDCCCLXVIII
EX LIBRIS

STUDIES IN EXPERIMENTAL EDUCATION

THE JOHNS HOPKINS UNIVERSITY
STUDIES IN EDUCATION

No. 3

EDITED BY
EDWARD F. BUCHNER
in cooperation with
C. MACFIE CAMPBELL

STUDIES IN EXPERIMENTAL EDUCATION

BY
BIRD T. BALDWIN, PH.D.
AND OTHERS

BALTIMORE
THE JOHNS HOPKINS PRESS
1920

PRESS OF
THE NEW ERA PRINTING COMPANY
LANCASTER, PA.

CONTENTS

CHARTS

TABLES

EDITOR'S NOTE

This number of the *Studies in Education* presents material collected and developed in an unexpected manner. The program of the 1917 Summer Courses of the Johns Hopkins University, under the writer's direction, included a Demonstration School of six grades, designed to be used by students for the observation of teaching in connection with several university courses in elementary education. It included grades *one* and *two,* taught by Miss Ida V. Flowers; *four,* by Miss Maude B. Smith; *five,* by Miss Helen M. Burnett; *six,* by Miss Matilda Srager; and *seven,* by Miss Julia F. Beck. One hundred fifty-five children were enrolled, the majority of whom were pupils in the Baltimore public schools who had failed of promotion in June, and who hoped by the six weeks' study to make their grades in September. Special teaching difficulties thus existed at the opening of the Demonstration School.

Instruction in Experimental Education (Education 1) was given in a university course by Professor Bird T. Baldwin, then of Swarthmore College, now Director of the Iowa Child Welfare Research Station. This course presented methods of educational measurements, with application to problems in the fields of physical growth, testing the growth of general intelligence, and the degrees of attainment by pupils in various school subjects. The class consisted of twenty-three students of graduate or senior college standing who were taking the course for the first time.

A belief in the special value of the results of educational measurements to the instructional problems of the grade teacher brought the instructor and students of Education 1 and the teachers and pupils in the Demonstration School together as investigators in a laboratory. The teachers of the grades, especially *four* to *seven,* were confronted with

the varying needs of the children which were to be met as fully as possible during the six weeks the school was in session. By measuring and testing the children, many of their individual needs were defined, and the information turned over promptly to the teachers for the guidance of their instruction. A large coöperative enterprise thus ensued during the session, to which many persons, both instructors and students, contributed.

While coöperating in the realization of the teaching aims of the Demonstration School, Professor Baldwin succeeded admirably in showing how a university summer course in experimental education can be organized so as to advance beyond instruction to investigation. Members of his class were assigned, by groups and individually, to problems in accordance with their special interests and previous training. He directed the giving of the tests and the formulation of the results, while his students are responsible for the details, order, and preservation of the data and the conclusions of their individual studies. The correlations have been most carefully checked up, and the findings are believed to be accurate. Professor Baldwin's early entrance upon Government service during the war greatly delayed the publication of these results. It is hoped that the selected twelve studies presented herewith will offer material and findings of special comparative value, and give additional impetus to the experimental movement in education.

The realization of the original teaching aims of the Demonstration School was largely due to the valuable assistance of Miss Florence E. Bamberger. The manuscript has been read in its entirety by her and by Dr. Buford J. Johnson, and in part by Capt. Richard M. Elliott. Special assistance has been rendered by Miss Agnes Snyder and Mr. William R. Flowers.

EDWARD F. BUCHNER.

I

INTRODUCTION AND SUMMARY

Bird T. Baldwin.

The papers assembled in this *Study* give a diagnostic picture of 129 out-of-step pupils who represent, in most instances, examples of maladjustment in educational progress. Modern experimental education orientates from the physical and mental development of the children who are being taught, as well as from the subject matter of instruction, and the results of these studies show the wide range of individual differences among children and the limitations of educational procedure designed to meet these differences.

In order to understand the conditions which aid or hinder education, it is essential that teachers with scientific, professional training should focus their attention upon the school-room situation, and analyze it into its various aspects. This, in short, has been the aim of this brief preliminary introduction to experimental education. The results are limited to the data at hand and no attempt has been made to formulate general conclusions beyond the results included in the study.

Value would have been added to the monograph if the observations had been continued throughout a year or a series of years, instead of a few weeks, and it is hoped that this work will encourage students to undertake such investigations. It is expected that trained experimenters will soon give us a more complete study on a similar basis, if education is going to become a science with experimental aspects. These may in a limited sense serve as type studies of a suggestive nature. They however have been prepared and published essentially for the benefit of the students who made them, in order that the students may have the advan-

tage of the unity of the course and the opportunity to improve on their own work. Publication has been delayed on account of the recent war.

A. Physical Measurements

The 129 children included in this school are, on the average, an inferior group when compared with the writers' norms obtained from consecutive measurements of children who have had continuous school-medical inspection, physical training and directed play. These tentative norms are given in Charts I to IV in order that comparisons may be made. The physical status of the Demonstration School children may be found on pages 22–26 of Miss Campbell's Report.

B. Mental Measurements

The term "mental age" is a gross blanket statement, since mental traits, abilities, interests and psychomotor reactions are not found equally developed in the so-called normal children of the same "mental age." That is, fundamentally, a scale graduated into groups or steps of "mental ages" is a rough approximation of what a number of children average, not what they are.

The measuring scales for intelligence, like those for subject matter, represent tentative approximations and not final fixed units. If all so-called normal children or a large percentage of them passed these particular tests we would have to raise our norm by increasing the difficulty of the tests.

The author of the Stanford Revision of the Binet Test states that the five or six tests that represent the "mental age" of seven, for example, are the five that 50 per cent of a supposed group of normal children can pass. The other 50 per cent might pass six other tests equally well. The point is that these six tests do not make a mental age as many are prone to think. It is the child that is normal and not the six tests. The tests are devices which catch certain combinations of traits, and "mental levels" are generaliza-

Name..... Schools.....

Date of birth..... Nationality..... Examiners.....

Date of examination													
School year	Kg.	1		2		3		4		5		6	
Age	5	6	6	7	7	8	8	9	9	10	10	11	11
Norm in inches	43.0	45.4	46.6	47.9	48.3	49.5	50.1	51.4	52.2	53.6	54.8	55.3	56.7
Height													
Weight-height coefficient	.95	.95	.96	.98	1.07	1.09	1.12	1.18	1.19	1.25	1.30	1.32	1 33
Norm in pounds	41.0	42.9	44.8	46.9	51.7	54.2	55.9	60.4	62.1	66.9	71.4	73.2	75.5
Weight													
Norm in cubic inches	50.0	65.6	70.0	77.2	82.4	85.6	92.6	95.7	105.9	111.5	123.2	121.6	126.2
Breathing capacity													
Vital-height coefficient	1.16	1.44	1.50	1.61	1.71	1.73	1.85	1.86	2.03	2.08	2.25	2.20	2.23

Date of examination													
School year	7		8		9		10		11		12		
Age	12	12	13	13	14	14	15	15	16	16	17	17	
Norm in inches	57.1	57.8	58.2	59.5	61.0	62.2	63.7	65.1	66.3	67.4	67.5	67.5	
Height													
Weight-height coefficient	1.42	1.46	1.54	1.51	1.56	1.59	1.65	1.78	1.79	1.89	1.90	1.88	
Norm in pounds	81.0	84.6	89.5	90.0	94.9	98.9	104.9	115.9	118.9	127.1	128.5	127.0	
Weight													
Norm in cubic inches	134.0	140.5	150.8	152.9	162.5	173.8	191.7	210.0	212.6	222.6	236.4	238.0	
Breathing capacity													
Vital-height coefficient	2.35	2.43	2.59	2.57	2.66	2.79	3.01	3.23	3.21	3.30	3 50	3.53	

These norms represent well-developed children with school-medical inspection and physical training. A small child for a given age may be well developed if the *coefficients*, height, weight and breathing capacity relationships, are normal and approximate those indicated. The formula is

Weight ÷ height = weight-height coefficient.
Breathing capacity ÷ height = vital-height coefficient.
For example. For 5 years, 41 lbs. ÷ 43 in. = .95 weight-height coefficient.
In the metric system use height in centimeters, weight in kilograms, and breathing capacity in liters. The English weight-height coefficient norms × .179 = metric weight-height coefficient norms. The English vital-height coefficient norms × .00645 = metric vital-height coefficient norms.

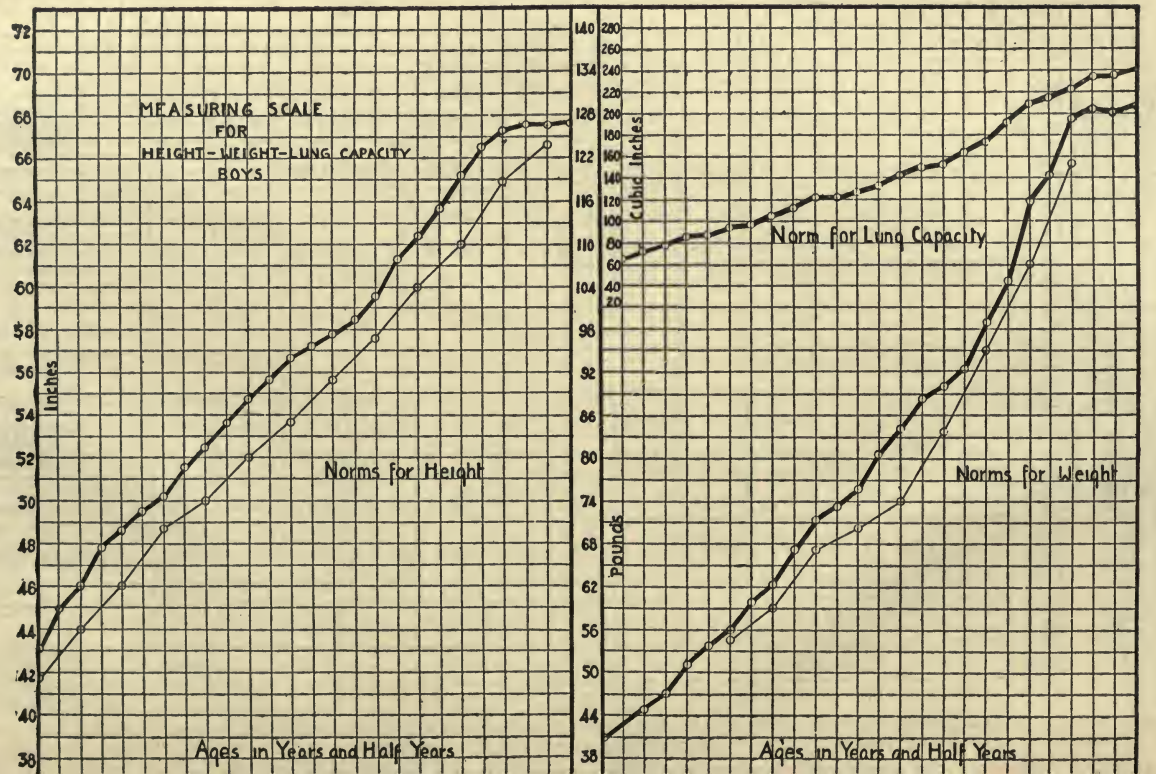
School year..... 1 2 3 4 5 6 7 8 9 10 11 12 | 1 2 3 4 5 6 7 8 9 10 11 12 | 1 2 3 4 5 6 7 8 9 10 11 12

Health..... Vision..... Hearing.....

Vaccination..... Glands..... Posture.....

Note: G = Good M = Medium P = Poor C = Corrected.

CHART II—REVERSE



MEASURING SCALE FOR PHYSICAL DEVELOPMENT (GIRLS)

English System

Name..... Schools.....

Date of Birth.....Nationality..... Examiners.....

Date of examination													
School year	Kg.	1		2		3		4		5		6	
Age	5a	6	6a	7	7a	8	8a	9	9a	10	10a	11	11a
Norm in inches	42 9	45.2	46.3	47.2	48.3	49.4	50 4	51.1	52.0	53.3	54.0	55.2	56.9
Height													
Weight-height coefficient	.93	.92	.95	1 02	1.04	1.11	1.13	1.13	1.18	1.22	1.27	1.30	1.32
Norm in pounds	40 1	41.7	44 2	48.1	50.4	54.7	57.1	57.5	61.5	64.8	68.4	71.7	74 9
Weight													
Norm in cubic inches	40.0	60 0	63.0	71.3	77.9	81.8	83.5	87.3	91.7	100 9	106 7	114.0	119.0
Breathing capacity													
Vital-height coefficient	.93	1.33	1.36	1.51	1.61	1.66	1.66	1.71	1.76	1 89	1.98	2.07	2.09

Date of examination													
School year	7		8		9		10		11		12		
Age	12	12a	13	13a	14	14a	15	15a	16	16a	17	17a	
Norm in inches	57.7	59.3	59.5	60.5	61.3	62.5	62.6	63.3	63.5	63.7	63.4	63.9	
Height													
Weight-height coefficient	1.44	1.47	1.54	1.59	1.67	1.69	1.76	1.82	1 82	1.84	1 86	1 86	
Norm in pounds	82.8	87.1	91.8	96 4	102.3	105.8	110.3	115.3	115.6	116.9	117 9	119.0	
Weight													
Norm in cubic inches	128.0	141 3	146.2	147.7	153.5	158 8	164.5	172.6	176.5	178.8	183 0	185.0	
Breathing capacity													
Vital-height coefficient	2.22	2.38	2.46	2.44	2.50	2.54	2 63	2.73	2 78	2.81	2 89	2.90	

These norms represent well-developed children with school-medical inspection and physical training. A small child for a given age may be well developed if the *coefficients*, height, weight, and breathing capacity relationships, are normal and approximate those indicated. The formula is

Weight ÷ height = weight-height coefficient.

Breathing capacity ÷ height = vital-height coefficient.

For example. For 5a years, 41 lbs ÷ 43 in. = .95 = weight-height coefficient.

In the metric system use height in centimeters, weight in kilograms, and breathing capacity in liters. The English weight-height coefficient norms × .179 = metric weight-height coefficient norms. The English vital-height coefficient norms × .00645 = metric vital-height coefficient norms.

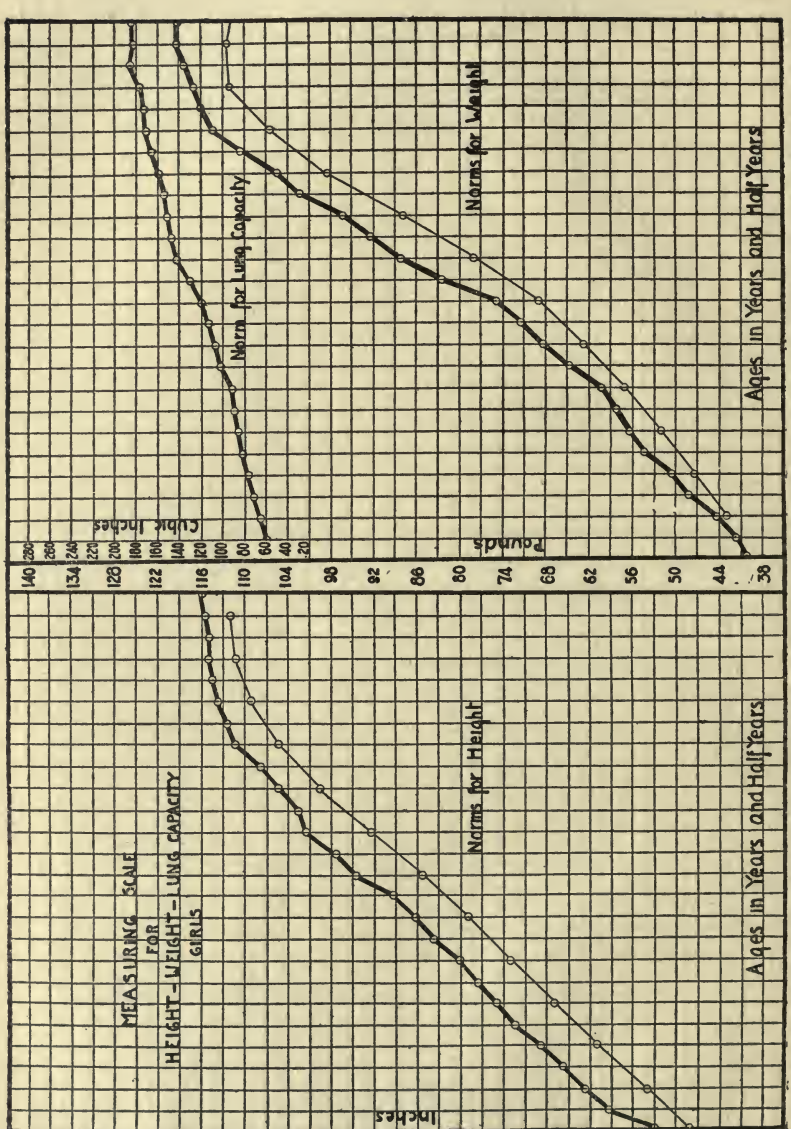
School year.....1 2 3 4 5 6 7 8 9 10 11 12 | 1 2 3 4 5 6 7 8 9 10 11 12 | 1 2 3 4 5 6 7 8 9 10 11 12

Health..... | Vision..... | Hearing.....

Vaccination..... | Glands..... | Posture.....

Note: G = Good M = Medium P = Poor C = Corrected

CHART IV—REVERSE



tions. The scales are too gross in their present forms for careful psychological diagnosis.

That a "mental age" is not a cross section of a combination of traits which fall within an average but a range of traits and a combination of traits may be shown experimentally. A "point scale" allows for individual differences and has psychological advantages over a step scale. A "point scale" should allow for a finer gradation of points than the Yerkes Scale and a wider range of tests if a general analysis of intelligence is the purpose. The new Yerkes Adolescent Scale offers some advantages here.

It will be seen that the Binet-Simon-Goddard-Terman scales are based fundamentally on a different point of view of mental development than the Huey-Yerkes-Bridges Point Scales. The one assumes that in the normal child the mind develops in pronounced stages or nodes and these nodes correspond in the main with certain chronological ages; the other is based on the presupposition that some traits may or may not develop before others, any or all may develop gradually—but the scale gives the credit for what is found, not what is supposed to be present at a given age.

The scales have been and are serving a good purpose if we remember that we are psychologists who see the child beyond the scale. The gathering together of the tests, the attempt to formulate norms and the practical use of these norms have all been worth while. In the future there will probably be a return to the laboratory type of psychological experiments where individual, specific traits are studied intensively as before, but with a much wider insight into their meaning which heretofore has been seldom considered and little understood.

The writer sees need in the future of the differentiation and refining of the scales into a series of graduated tests for each trait or ability. It will be possible by a combination of scores to make up a norm. For example, there should be a graduated series for auditory memory of digits, for words, for sentences, or pictures, etc., and also for motor control, for different types of judgment, and so on.

TABLE 1

SHOWING THE DISTRIBUTION AND OVERLAPPING OF THE MENTAL TRAITS OF 129 PUPILS IN THE DEMONSTRATION SCHOOL AT JOHNS HOPKINS UNIVERSITY

Points	Mental Ages												Test No.
	4	6	7	8	9	10	11	12	13	14	15	16	
3	1	3	1	9	6	6	21	10	9	27	28	9	I
2		1								1			
1	1												
0													
4		1		7	5	5	20	9	9	28	27	9	II
3				1	1		1				1		
2		1	1	1				1					
1		2				1							
0	2												
3		1	1	8	4	5	20	9	8	28	27	9	III
2							1	1	1				
1	1	3		1	2	1					1		
0	1												
5				1	2	3	5	7	3	15	18	5	IV
4		2		4	1		6	2	4	10	8	3	
3	1	1	1	3	3	3	10	1	1	3	1	1	
2	1	1		1					1		1		
1													
0													
4	1			5	6	6	21	10	9	28	28	9	V
3		2	1										
2				2									
1				1									
0	1	2		1									
6				1	1		1	1	2	11	5	7	VI
4				1	1	3	7	7	4	12	17	2	
2	1	4	1	7	4	3	13	2	3	5	6		
1	1												
0													

Points	Mental Ages												Test No.
	4	6	7	8	9	10	11	12	13	14	15	16	
9						1	5	3	2	14	18	7	VII
8						2	2		3	4	4	1	
7			1	2	3		6	3		4	3	1	
6				3	2	3	8	4	4	5	3		
5		2		3	1					1			
4													
3	1	2		1									
2													
1													
0	1												
2				4	3	5	14	7	5	23	25	9	VIII
1	1	2		2	1	1	3	2	2	4	2		
0	1	2	1	3	2	2	4	1	2	1	1		
6				1	3	1	12	3	5	17	23	8	IX
5				1	2		4	4	2	6	3	1	
4		1		1		2	3	2	1	5	2		
3		1		2		1	1	1					
2		1		4		1	1		1				
1			1		1								
0	2												
8						1		1	3	8	12	6	X
7					1				2		5	3	
6					1	1	1	1	1	6	3		
5		2		2	1	1	1	7	2	7	5		
4			1	3	2	2	2	1		4	2		
3		1		1	2	2	2		1	2	1		
2		1		2									
1													
0	2			1									

Points	Mental Ages												Test No.
	4	6	7	8	9	10	11	12	13	14	15	16	
3				6	1	4	12	6	7	17	21	5	XI
2	1	1		1	3	1	4	3		6	3	3	
1		2		2	2		5		2	4	3	1	
0	1	1	1			1		1		1	1		
4		3		6	3	3	18	8	6	22	25	9	XII
3				3	1	3	1	1	2	4	2		
2	1			1	1		1	1	1	1	1		
1	1	1	1		1		1			1			
0							1						
4						2	2	1	4	14	14	8	XIII
3				2	1	1	6	4	4	6	7	1	
2			1	3	1	3	9	3		6	5		
1		1		1	1		1	1	1	2	1		
0	2	3		3	3		3	1			1		
4						3	8	7	3	17	18	8	XIV
2				1		2	8	3	3	9	9		
0	2	4	1	8	6	1	5		3	2	1	1	
8						1	2	2	1	10	18	7	XV
7										1	1	1	
6							5	4	1	10	7		
5								1		3	2		
4				1	4	5	12	3	4	2			
3					1				3	1			
2		2	1	5	1		2			1			
1		1											
0	2	1		3									

Points	Mental Ages												Test No.
	4	6	7	8	9	10	11	12	13	14	15	16	
4				1	1		5	5	5	14	18	7	XVI
3				2	2		4		1	1	6	2	
2				3	2	1	7	4	2	7	2		
1		1	1	3	2	1	2	1	1	3	2		
0	2	3		3	1	4	3			3			
5							1	5	1	8	17	7	XVII
4					1		5	2	5	12	8	1	
3				2	2	2	7	2	2	4	1	1	
2						2	6	2	1	2	2		
1					1	1	1	1		2			
0	2	4	1	7	2	1	1						
6							4	2	4	10	14	7	XVIII
4					1	1	8	5	2	12	9	2	
2					3	2	6	2	1	3	5		
0	2	4	1	9	2	3	3	1	2	3			
6							2	2	1	10	18	7	XIX
4					1	2	8	3	5	12	7	2	
2				3	2	3	9	1	2	5	1		
0	2	4	1	6	3	1	2	4	1	1	2		
6									1	4	4	4	XX
5						2	1	1	4	4	3	2	
4						1	1	2		7	8	2	
3					1		6	1	3	6	6	3	
2				4	3	2	8	3	1	4	7	1	
1					1	1	3	2		2		1	
0	2	4	1	5	1		2	1		1			

Normal children differ greatly in mental characteristics, and it is this range of individual differences that makes up the class which we should call normal children, and not an average of them. It is the average of this child's own traits that determines whether or not he is normal. This may be proven experimentally by the examination of the records of the 129 children. These pupils have been given the Point Scale tests as shown in Table 1. If on the basis of these tests, the individuals are grouped according to "mental ages," it is found that there is a remarkable *overlapping* of abilities. For example, tests that are designed by Binet, Goddard and Terman, for 7 to 11 years of age are often very poorly executed by children that mentally score ages from 12 to 16. To be more specific, take the test for drawing the two designs placed by Binet for 11 years and by Terman for 10 years of age. In Table 1 are a large number of children that are 12, 13, 14, 15, 16 or 17 years old mentally according to the rating of the mental tests that do not make a passing score on this test. The same is true of every other test in any series except the very easy ones.

This overlapping in abilities is very evident in an unpublished investigation made by the writer of 1,500 delinquent children.

C. Correlations

In the study of psycho-educational problems and processes, it is frequently desirable to measure the relationship which may exist between two series of observations, tests, or measurements. It is most difficult, if not impossible, to estimate this relationship by simply observing the two series. The relationship may conveniently be expressed by means of a single numerical expression or coefficient of correlation. Probably the most satisfactory coefficient is that devised by Karl Pearson. The formula is $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$ where x is the deviation from the arithmetic average (signs considered) for one series and y is the deviation from the

arithmetic average for the other series. The series are arranged so that the corresponding items are opposite each other and the product xy is for the deviation of the corresponding items. For example, in correlating the results of two measuring scales, the amount which a child's record deviates from the average record in one test, is multiplied by the amount which the same child's record deviates from the average in the other test. The sum is taken algebraically. If the two series correspond exactly in their deviations in the same direction, the result is complete or positive correlation, or + 1. If the two series correspond exactly in their variations, but in opposite directions, the result is completely negative correlation or — 1. If no relation exists between the two series, the degree of correlation is 0. Intermediate values may exist anywhere between — 1 and + 1. The probable error of a coefficient has been carefully worked out and may be obtained by the formula $\frac{.6745(1-r^2)}{\sqrt{n}}$, where n is the number of pairs of items. It indicates that the coefficient actually lies between r plus the probable error and r minus the probable error. The size of the probable error always varies inversely with the size of the coefficient and with the number of items. The coefficient should be over .30 to show correlation, and .50 or over indicates decided correlation if the coefficient is at least six times the probable error.

In order to determine whether or not the indices of intellectual ability as measured by the Yerkes-Bridges scale in this particular group of children showed any direct relationship to the physical measurements of height, weight, grip and breathing capacity, coefficients of correlation were determined by means of the Pearson formula for all measurements of 67 boys ranging from 6 to 16 years of age, and 60 girls ranging from 5 to 16 years of age.

The results show no correlation for the boys, since the coefficient for intellectual ability and height is — .228; for in-

tellectual ability and weight —.135; for intellectual ability and grip —.226; and for intellectual ability and lung capacity —.197 with the probable errors ±.078, ±.081, ±.078, and ± .079, respectively. In other words with this group of children there is no evidence that physical growth as indicated by height, weight, grip, and breathing capacity shows a positive or negative relationship with intellectual ability as indicated by the Yerkes-Bridges Scale for measuring intelligence. For the I. Q. (Stanford), the coefficients are —.351, —.287, —.293, —.314, respectively. Here there is a slight negative correlation.

For the 60 girls, results indicate very slight correlation with the tendency toward the negative direction. For intellectual ability and height the coefficient is —.307, with the probable error of ± .079; for intellectual ability and weight —.166 with a probable error of ± .085; for intellectual ability and grip —.190 with a probable error of ± .084; for intellectual ability and breathing capacity —.069 with a probable error of ± .087. For the I. Q. (Stanford), the coefficients are —.449, —.377, —.397, —.343, respectively. Here there is more marked negative correlation.

In interpreting results from these data, it must be born in mind that this group represents both retarded and accelerated pupils, that while many children came to school because they were below grade, others came to take advantage of the opportunity for promotion which would follow the completion of one or more subjects. The girls are inferior to the boys, mentally.

Since these same children had been tested by thirteen scales designed to measure degree of attainment in subject matter, the relationship between the results of these tests and intellectual ability was determined. The coefficient of correlation for each scale and intellectual rating as measured by the Stanford Revision Scale and by the Yerkes Bridges Scale was obtained for boys and girls separately. The coefficients with their probable errors are given in Table 2.

TABLE 2

COEFFICIENTS OF CORRELATION FOR PER CENTS OF INTELLECTUAL ABILITY FOR THE YERKES-BRIDGES SCALE FOR MEASURING INTELLIGENCE AND INTELLIGENCE QUOTIENTS FOR THE STANFORD REVISION SCALE FOR MEASURING INTELLIGENCE WITH THIRTEEN SCALES FOR MEASURING DEGREES OF ATTAINMENT IN SUBJECT MATTER*

Measuring Scales	Stanford Intelligence Quotients						Yerkes Coefficients for Intellectual Ability					
	Boys			Girls			Boys			Girls		
	Coef.	P. E.	No.	Coef.	P. E.	No.	Coef.	P. E.	No.	Coef.	P. E.	No.
Woody Addition	+.075	±.090	55	+.185	±.091	52	+.058	±.091	55	+.305	±.085	52
Woody Subtraction	−.069	±.090	56	+.034	±.098	47	+.045	±.090	56	+.260	±.091	47
Woody Multiplication	+.024	±.089	58	+.030	±.092	54	+.004	±.089	58	+.257	±.086	54
Woody Division	+.010	±.089	58	+.067	±.092	54	+.056	±.089	58	+.299	±.084	54
Ayres' Handwriting	−.215	±.085	58	+.044	±.092	54	−.086	±.088	58	+.189	±.089	54
Freeman Handwriting	−.148	±.087	57	+.051	±.092	54	+.020	±.089	57	+.203	±.088	54
Thorndike Handwriting	−.115	±.088	57	+.048	±.092	54	−.027	±.089	57	+.184	±.089	54
Kansas Silent Reading	+.437	±.072	58	+.483	±.071	54	+.341	±.079	58	+.434	±.075	54
Starch Comprehension	+.210	±.087	55	+.569	±.065	49	+.143	±.089	55	+.474	±.074	49
Ayres' Spelling	+.160	±.087	57	+.137	±.091	53	+.099	±.088	57	+.351	±.082	53
Starch Spelling	+.218	±.086	56	+.127	±.092	53	+.220	±.086	56	+.320	±.093	53
Completion Scale	+.213	±.085	58	+.177	±.090	53	+.219	±.085	58	+.361	±.081	53
Composition	−.121	±.089	56	+.014	±.094	52	+.009	±.090	56	+.123	±.093	52

* The coefficient of correlation between the Stanford Scale and the Yerkes-Bridges is +.708 for the boys, and +.806 for the girls.

This table shows the fallacy of basing promotions primarily on intelligence ratings as a number of psychologists are advocating.

As a whole there is little indication of correlation between intellectual rating and the different measuring scales. In the case of the three handwriting scales there is no correlation for either boys or girls between handwriting ability and the intelligence quotients or intellectual ability, since the coefficients vary between —.215 and +.203. Also, there is no correlation between either the intelligence quotients or intellectual ability and the Woody tests for abilities in addition, subtraction, multiplication, and division, the coefficients being very close to zero except for the girls' coefficients for I. A. and there the highest coefficient is +305 with a probable error of ±.085. With the English composition scale there is no evidence of correlation and for the Trabue completion scale the coefficient is low. There apparently is slight correlation between intellectual ability for the girls and Ayres' spelling, since the coefficient is +.351, but this can be given little meaning, since the correlation for the boys is only +.099 and for the intelligence quotients with boys and girls +.160 and +.137 respectively. The coefficients of correlation between Starch's spelling and the Stanford intelligence quotients are positive but low, ranging from +.127 to +.320. With Starch's Comprehension Scale and the Kansas Silent Reading Test, a positive correlation is found though it is not high. The correlation between the intelligence quotients and the intellectual ability for the girls and their records in the Starch comprehension scale is good, for the coefficients are +.569 and +.474 respectively with probably errors of ±.065 and ±.074. For the boys, however, there is little or no correlation. For them, the coefficients are +.437 for the Kansas Silent Reading Test and the intelligence quotients (I. Q.), with +.341 for the Kansas Silent Reading Test and intellectual ability (I. A.); for the girls, the coefficients for the Kansas Silent Reading Test and the intelligence quotients (I. Q.) is +.483, and for the Kansas Silent Reading Test and intellectual ability (I. A.) +.434. This test is, therefore, more closely correlated with general intelligence than any other measuring scale included in this list.

D. Measuring Scales in Subject Matter

A number of inter-correlations for the measuring scales were computed and these coefficients with their probable errors are given in Table 3. In writing, the Ayres, Thorndike, and Freeman Scales were correlated each with each

TABLE 3

Inter-Correlations for Scales Measuring Degree of Attainment in Subject Matter

Correlation Between	Boys			Girls		
	Coef.	P. E.	No.	Coef.	P. E.	No.
Woody Addition and Woody Subtraction	+.814	±.031	54	+.776	±.040	45
Woody Addition and Woody Multiplication	+.748	±.040	55	+.774	±.038	52
Woody Addition and Woody Division	+.776	±.036	55	+.697	±.048	52
Woody Subtraction and Woody Multiplication	+.763	±.038	56	+.853	±.027	45
Woody Subtraction and Woody Division	+.812	±.031	56	+.780	±.038	47
Woody Multiplication and Woody Division	+.795	±.033	58	+.805	±.032	54
Ayres Handwriting and Thorndike Handwriting	+.915	±.014	57	+.612	±.057	54
Ayres Handwriting and Freeman Handwriting	+.903	±.016	57	+.593	±.059	54
Thorndike Handwriting and Freeman Handwriting	+.875	±.021	57	+.396	±.077	54
Trabue Completion Scale and Kansas Silent Reading Test	+.217	±.084	58	+.301	±.085	53
Starch Comprehension and Composition	+.116	±.090	54	+.264	±.090	48
Ayres Spelling and Starch Spelling	+.824	±.029	56	+.756	±.039	53

and the results thus obtained show very high correlation for the boys, the coefficients being +.915, and +.903, +.875. In the case of the girls, there is positive but not very high correlation, the coefficients being +.612, +.593, +.396. This may indicate that the handwriting scales are graded more in accordance with boys' writing than with girls' writ-

ing. In Arithmetic the Woody tests for addition, subtraction, multiplication, and division were correlated each with the other, and in every case the coefficient shows a decided positive correlation, equally high for boys and girls. The highest coefficient is +.853 with a probable error of ±.027 and the lowest is +.697 with a probable error of ±.048. In spelling very high correlation between the Ayres and Starch scales is shown by the coefficients +.824 and +.756 for boys and girls respectively. The Starch Comprehension Scale when correlated with the Composition Scale gives coefficients of +.116 and +.264 for the boys and girls, or no correlation. There is little correlation between the Trabue Completion Scale and the Kansas Silent Reading Test, since the coefficients for the boys and girls are only +.217 and +.301. Therefore, all the high correlations are between very closely related tests, such as the three handwriting scales, the four Woody Arithmetic tests and the two spelling scales.

E. A Developmental Graph of the Traits of an Accelerated Boy and a Retarded Boy

This circular diagram is designed to show relatively and comparatively the individual degrees in attainment in physical measurements, mental tests and various scales for measuring ability in subject matter. A group of 60 boys and another of 55 girls were measured and tested individually in twenty-two physical and mental traits. The group was composed of pupils of the 4th to 7th grades inclusive. The average score for the whole group was obtained in each trait or test and this value is given in terms of 100 per cent and posited as a working norm for comparative ratings. Diagrammatically, this is represented by the heavy circle with radii representing 100 per cent accomplishment. The individual's rating for any one test is divided by the group average for that test which gives the per cent of this average score for the individual. This is indicated on the diagram by the length of a radius. For example, the average chrono-

logical age for this group of boys is 12 years and 8 months. The accelerated boy in the diagram was only 11 years and 3 months, or 89 per cent of the average age. But his mental age as scored by the Stanford Revision Scale was 14, with an average for the group of 12 years and 2 months. Therefore, his mental age by the Stanford Revision Scale is rep-

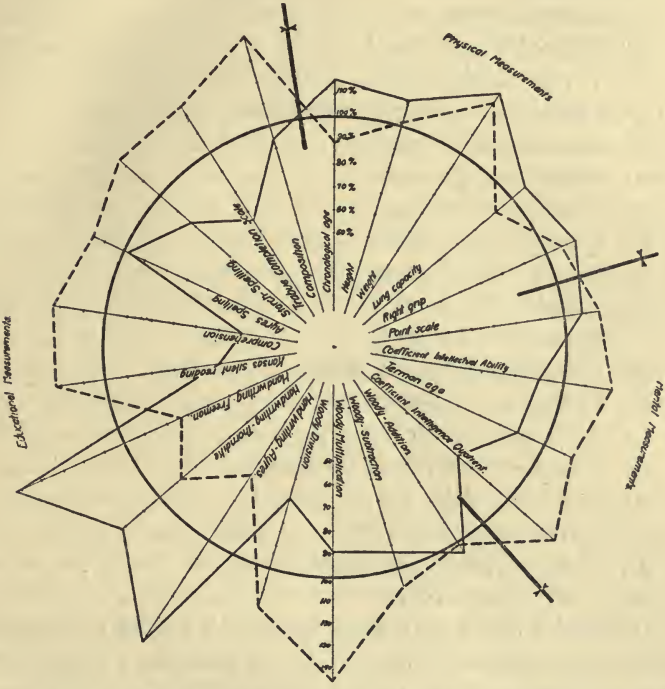


CHART V.—A Developmental Graph Showing Per Cent of Attainment in Terms of Group Averages

Group Average, or Norm.
Accelerated Boy.
Retarded Boy.

resented on the circle by 115 per cent. His mental age by the Yerkes Bridges Point Scale is 16 or 116 per cent of the average Point Scale age of 13 years and 9 months. In every case, the length of the radius indicates the per cent of the average, and the results in all tests are thus made comparable.

The diagram shows two individuals records selected from the complete series of 115, an accelerated boy and a retarded boy of approximately the same chronological ages, for 22 traits including,

1. Chronological Age
2. Standing Height
3. Weight
4. Lung Capacity
5. Grip of Right Hand
6. Yerkes Bridges Point Scale
7. Coefficient of Intellectual Ability
8. Stanford Revision Age
9. Intellectual Quotient
10. Woody's Addition
11. Woody's Subtraction
12. Woody's Multiplication
13. Woody's Division
14. Ayres' Handwriting
15. Thorndike's Handwriting
16. Freeman's Handwriting
17. Kansas Silent Reading Test
18. Starch's Comprehension Test
19. Ayres' Spelling Test
20. Starch's Spelling Test
21. Trabue Completion Scale
22. Composition Test

It will be noted that there are wide ranges of degrees of accomplishment in each trait for each boy; in physical measurements, and in writing the retarded boy excels the accelerated pupil. In the other traits the accelerated pupil is decidedly superior to the retarded boy.

It should be noted that these standards, norms, scales and tests represent tentative approximations and not fixed nor final units. Some have been obtained through the consensus of opinion of educators and school men and represent an "average" point of view, as is illustrated by a few of the English Scales. Some represent the average or median at-

tainments of pupils in representative schools, in certain aspects of a subject; of these the Courtis or Woody tests are examples. Some represent the results of a psychological and educational analysis of the perceptions involved in the learning process, such as the Freeman Scale in Writing. Some are types of psychological experiments which have long been used in laboratories, as Trabue's use of the Ebbinghaus Completion Test, while others represent inductive, consecutive studies on a limited number of selected children, as Baldwin's Growth Scores.

The fundamental point to be understood is that none of these standards are final or fixed and furthermore that they should not be. The educative process is a changing, progressing, developing process within the individual, and the same principle of growth and adjustment should hold within a school system from year to year and from generation to generation. All of these scales are more or less mechanical devices to help foster growth within the individual, within a school grade, and within a school system. All are means and not ends. Each of the scales gives cross section points of view from particular angles, and it may happen that two scales in the same subject will contradict each other, and yet each be correct and valuable for the particular purpose for which it was designed. The scales and tests are valuable as checks and guides; but they should not be set up as permanent goals nor even as immediate ends for all children, for education is an individual matter and the individual capacities vary widely. There is always a wide *range* of individuals within a class, but never an *average* individual who represents all of these differences. Teachers and parents must learn to think in terms of the *range of individual differences.* At the same time an effort should be made to bring the class as a whole up to a certain degree of attainment as represented by some particular *scale.* It is conceivable that a child might excel in one, two, or even more of the scales and still be inferior in other phases of the subject. This is demonstrated by the fact that there is much

overlapping in the results obtained by the lower and upper grades and also at the earlier and later ages.

The science of Experimental Educational Psychology is worthy of detailed consecutive study under standarized conditions which can be controlled, repeated, modified and compared. The next logical step is to determine *what traits and processes are being measured and to evaluate their significance in physical and mental development.*

II

PHYSICAL MEASUREMENTS

Laura Winder Campbell and Harry J. Kefauver

This investigation deals with the physical characteristics of one hundred and forty-one children in the Johns Hopkins Summer Demonstration School including grades I, II, IV, V, VI, and VII; and compares them with the norms for the corresponding chronological ages as established by former investigators.

The measurements taken were the standing height in stocking feet; sitting height; weight, from which two pounds and a half were deducted for clothing; grip tested three times alternately with each hand of which the highest measure was used for each hand; breathing capacity tested three times and highest results used. The stadiometer was used to measure the height, the hand dynomometer for the grip, the scales for the weight, the wet spirometer for the breathing capacity. The age of the nearest birthday was adopted. The data for each individual were arranged on a card so that in compiling results, the cards could be shuffled according to age, grade, height, weight, etc. The results were computed separately for each sex; the measurements were averaged for each age and also for each scholastic grade. Tables for each measurement are submitted showing the range of distribution, medians, and deviations for each age group. Also graphs have been plotted showing average measurements for each group in comparison with the Baldwin and Smedley norms for height, weight, and breathing capacity, and for right and left grip. The average instead of the median was taken, as the group of individuals at each age for each sex separately was small. In each of the graphs, the solid lines show the results of this

investigation while the broken lines represent the norms used.

On the whole it will be seen from the graphs that the measurements are below, though approximating the norms: the greatest divergence occurs in the case of the group of 14 year old girls where the measurements are lower than those of a normal child six months or a year younger.

Height

From the age-height distribution table, it is evident that at the ages of 11, 12 and 13, the tallest child is a girl while at all ages except 12 and 14, the shortest is a boy. At the ages of 10, 13, and 15, the girls exceed the boys in height. It is to be remembered in this connection that the 14 year

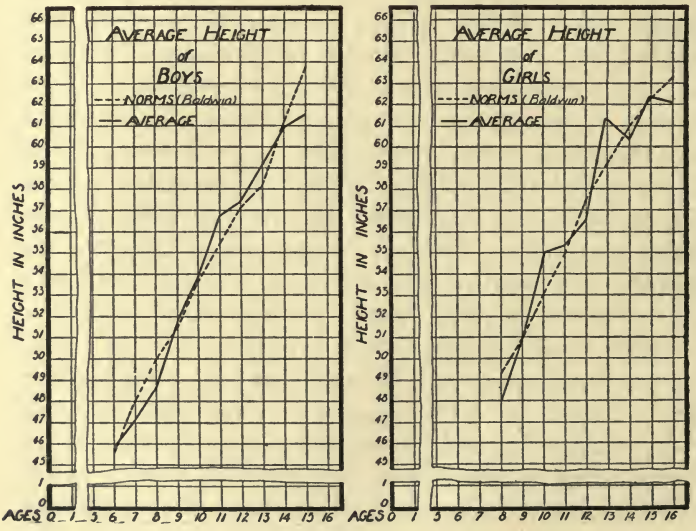


CHART VI

old group is shorter than the 13 year old group and shorter than the norms. The variation of individual heights in the different age groups is greatest for boys between the ages of 13 and 15; for girls from 11 to 14 years. These facts

are in accordance with the conclusion that girls reach their period of accelerated growth at an earlier age than boys.

The ratio between the total stature and sitting height varies with little regularity; in the case of the boys, this ratio increases in variable amounts from 6 to 8 years of age; decreases to 13 years and then increases. The girls show less irregularity; the ratio rises from 10 to 16 years (with the exception of the 14 year group) though with varying increments. This is not in accordance with the conclusion reached by Boas that this ratio decreases as age advances until 13 years of age for girls and 15 for boys when it increases in each case. In about 45 per cent of the individuals, the total height of one child was found to be several inches less than that of another while the length of trunk of the shorter individual was several inches greater than that of the taller. This variability was found to occur between the ages of 12 and 14 for both boys and girls, evidently the period of adolescence. The coefficient of correlation by the Pearson formula between the total stature and the sitting height was found to be +.959 for boys, and +.957 for girls.

Weight

The weight of the individuals of this group as measured in pounds extends over greater range than the height in inches. Ages 11, 13 ,14, and 15 show greatest difference between individuals of same age group for boys, and 12, 14, and 15 for girls. It is to be noted that at all ages the average weight of the girls range from two to fifteen pounds less than the normal with the single exception of the 13 year old group which surpasses the norms not only in weight but in the other measurements as well. The heaviest child at all ages except 12 and 15 is a boy while the lightest at 11, 12, and 14 is a girl. This does not show an exact parallel with the tallest and shortest of these age groups. The weight increases, as is the case with the height, most for the boys from 11 to 12 years of age; and from 12 to 13 for the girls. Again it must be recalled, however, that the 13 year old

group of girls exceeded the norms while the 12 year group was subnormal.

In plotting the weight-height indices for the age groups, they are found to be below the Baldwin norms for the boys

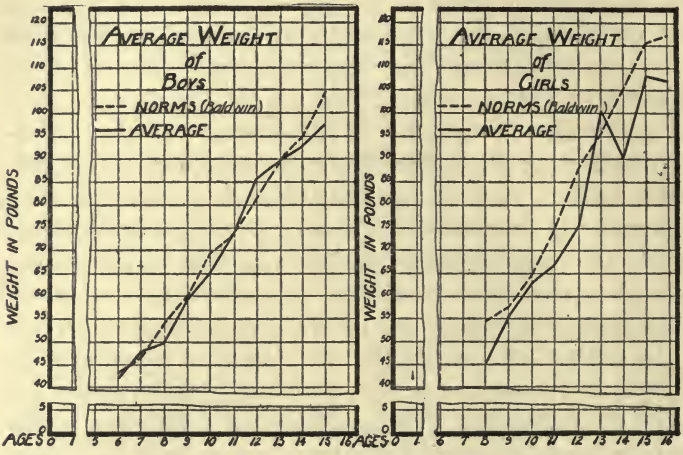


CHART VII

in each age group except at 12 years; this divergence was greater among the girls than among the boys, the exception occurring at the 13 year group which is higher than the norms.

Grip

While in the case of some individuals, the grip of the left hand was found to be greater than that of the right, the averages of the age groups showed the right hand to be the stronger. In the case of the thirteen year old boys, the grip of the right hand was the same as at the age of 12 years though the left hand grip increased a pound and a half from 12 to 13 years. Only in the 13 year group did the girls exceed the boys in this test, but this may be explained by the fact that a girl of 13 surpassed all of the other girls and boys but one in grip, thus raising the average of the 13 year group. The boys increased in grip more between the ages

of eight and nine years, and 11 and 12 years, than at any other age; and the girls showed the greater increase from 14 to 15 years. The 14 year old group of girls, however, was found to be sub-normal.

The correlations with height were + .911, and + .733 for boys and girls respectively.

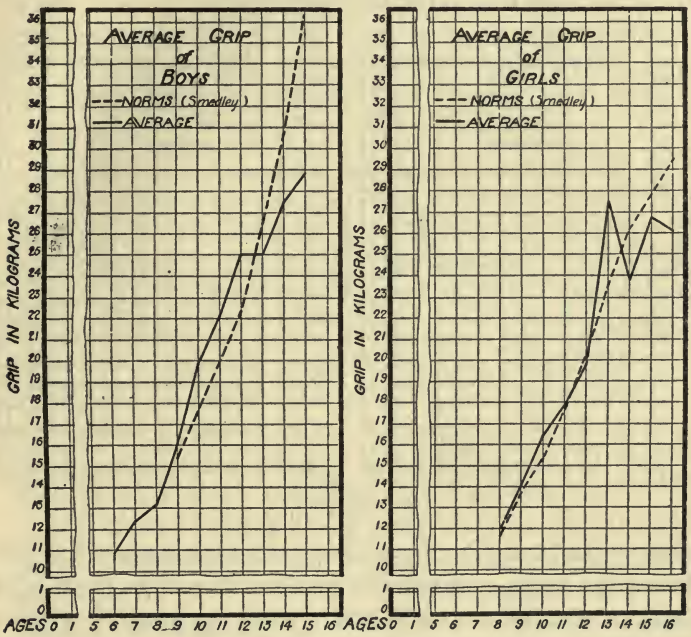


CHART VIII

The average age of each grade of this school was found to be from one to two and a half years greater than the normal age, the girls being more retarded than the boys. In the average measurements for each grade, the boys have been only slightly below the norms and in a few cases above. The girls, however, have been sub-normal in every case with the single exception of the height of the sixth grade.

Lung Capacity

On account of the small number of cases in certain particular ages we have given an age-grade distribution table with the average lung capacity for each year. In comparing these norms with the norms of Baldwin's investigation we find them as a general rule lower. Among the boys from five to ten years our norm follows the regular norm. Among the girls from five to ten the lung capacity is above the av-

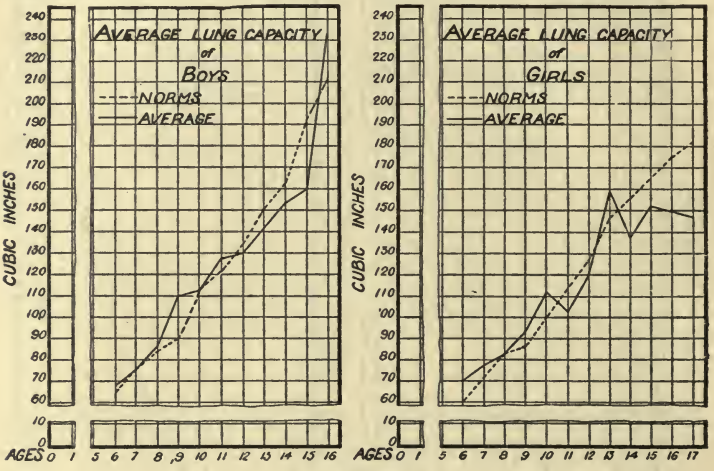


CHART IX

erage. At 11 and 12 the girls show a considerable decrease below the norm. Among the boys for these ages there is an increase for the 11th year and a decrease from the 12th to the 16th year inclusive. In fact all the ages among the boys are below the average. There is one case only in the 16th year, especially good, hence the reason for this variance.

In the 13th year for the girls there is a marked increase over the average, and from the 14th year to the 17th year, inclusive, there is a marked decrease in the average lung capacity. By an examination of the other measurements of the group of girls in the 13th year there is found a corresponding increase. This corresponds to the results of other investigations. Physically this is a retarded group of boys and girls.

III

THE APPLICATION OF THE YERKES-BRIDGES POINT SCALE AND THE STANFORD REVISION OF THE BINET SCALE FOR MEASURING INTELLIGENCE

Robert L. Bates, Nora V. Boston, S. M. Clark, W. R. Flowers, Susan Z. Housekeeper, Aimee Jones, Rosalie R. Martin, and Alice W. Ratcliff

In order that the reader may see the comparative results for all pupils tested by the Point Scale and the Stanford Revision or Terman Scale, the results are given below in tabulated form. In this table, G. signifies grade; C. age, chronological age; Pt. Sc., Point Scale; I. A., coefficient of intellectual ability; T. age, Terman age; I. Q., intelligence quotient. According to Stanford Scale all children whose I. Q. is 70 or below are considered mentally deficient. In this group, measured by the Point Scale, 1 boy and 3 girls are recorded as mentally deficient, but measured by the Stanford Revision Scale, 1 boy and 4 girls are mentally deficient.

The Yerkes-Bridges Point Scale is easier than the Stanford Revision of the Binet Scale after the "mental age" of eleven years, as may be seen by Table 4 where the individual scores are given. The Point Scale results are found in mental ages by comparing an individual's total score with expected scores in norms. The Coefficient of Intelligence is found by dividing the actual score by the expected score. The results for the Stanford Revision Scale are scored in terms of years and months and the Intelligence Quotient is found by dividing the mental age by the chronological age. The Coefficient of Intelligence and the Intelligence Quotient are not directly comparable and must not be confused.

In the Demonstration School, 129 children were given individually the Stanford Revision of the Binet Scale for rat-

ing intelligence and the Yerkes-Bridges Point Scale. The individual measures are given in Table 4. The general results of these examinations are:

		Stanford Revision	Yerkes-Bridges
Retarded	6 years	2	2
	5 years	1	1
	4 years	7	3
	3 years	8	5
	2 years	21	7
	1 year	26	13
Normal		35	30
Accelerated	1 year	15	30
	2 years	10	17
	3 years	3	5
	4 years	1	10
	5 years	0	3
	6 years	0	1
		Not tested by Yerkes Scale 2.	

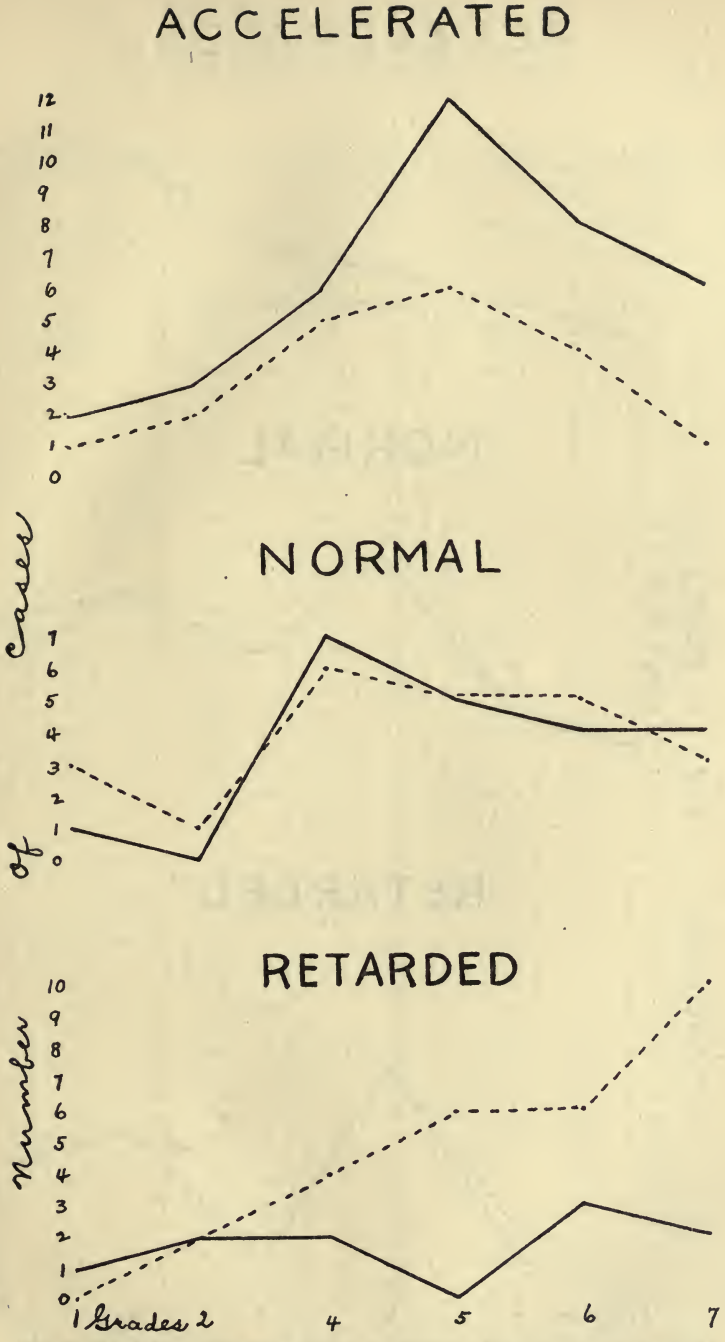


CHART X.*—A Comparison of the Two Scales in Estimating the Mental Ability of 70 Boys

Point Scale ———— Terman - - - - - -

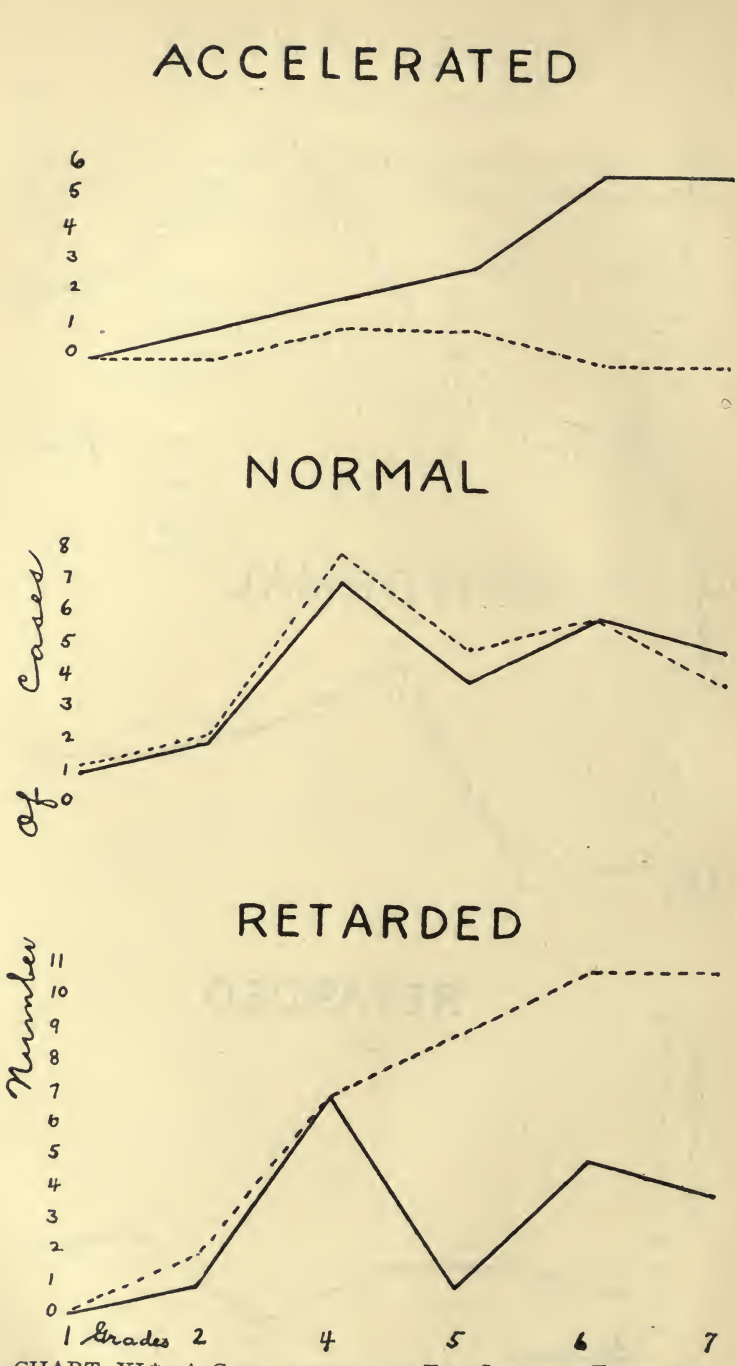


CHART XI.*—A Comparison of the Two Scales in Estimating the Mental Ability of 61 Girls

Point Scale ———— Terman ------

TABLE 4

OMPARISON OF MEASURES BY THE POINT SCALE AND THE STANFORD REVISION SCALE[1]

No.	Boys						No.	Girls					
	G.	C. Age	Pt. Sc.	I. A.	T. Age	I. Q.		G.	C. Age	Pt. Sc.	I. A.	T. Age	I. Q.
1	1	6	6	103	6–4	106	1	1	5–3	6	123	6–2	117
2	1	6–2	8	144	7–8	124	2	2	6–6	9	165	7–4	113
3	1	6–8	8	128	7–2	108	3	2	7–8	7	91	7–6	98
4	2	6–8	8	154	7–6	113	4	2	8–10	8	100	7–10	89
5	1	6–9	6	100	6–4	94	5	4	8–11	14	153	11–3	126
6	2	7–1	9	148	8–4	118	6	2	9–5	8	74	8–4	88
7	2	7–5	8	141	9–4	126	7	4	10–1	9	96	9–10	98
8	2	7–11	6	84	6–8	84	8	4	10–3	8	86	9–2	89
9	2	8	4	36	4–8	58	(9)	4	10–3	8	69	8–10	86
10	4	9–3	10	117	10–3	111	10	5	10–5	12	121	11–10	114
11	4	9–5	10	110	9–10	104	11	5	10–9	11	109	10–3	95
12	4	9–5	11	122	10–2	108	12	4	10–11	11	108	11–2	102
13	5	9–9	12	129	11–10	121	13	5	10–11	12	118	11–9	108
14	4	10–2	9	98	9–10	97	14	5	11–2	12	113	10–9	96
15	4	10–2	14	132	11–3	111	15	4	11–4	11	103	11–9	104
16	5	10–2	15	146	12–7	124	16	4	11–4	9	84	9	79
17	5	10–3	14	130	11–4	111	17	6	11–4	15	126	10–10	96
18	4	10–4	12	124	11–9	114	18	5	11–5	15	130	11–8	102
19	4	10–8	14	126	12–3	115	19	4	11–6	10	98	10–8	93
20	5	10–11	15	134	11–8	107	20	4	11–7	11	98	11–3	97
21	5	10–11	15	134	12–4	113	(21)	6	11–8	13	108	12–3	105
22	4	11	11	106	9–11	90	22	4	11–10	11	94	11–4	96
23	5	11	11	101	10–10	98	23	4	12–2	14	114	12–1	99
24	4	11–1	11	104	10–10	98	24	6	12–2	14	105	12–4	101
25	4	11–1	12	118	10–10	97	25	6	12–2	14	112	11–10	97
26	6	11–2	14	131	13–6	121	26	6	12–4	13	102	11–10	96
27	4	11–3	11	103	9–2	81	27	6	12–6	11	90	10–8	85
28	5	11–3	10	95	10–1	90	28	4	12–8	10	82	9–4	74
29	5	11–3	16	134	14	124	29	4	12–8	11	89	11–1	88
30	6	11–3	Adult	148	15–2	135	30	6	12–8	15	113	11–11	94
31	4	11–4	15	129	13–6	119	31	7	12–8	15	113	13–5	106
32	4	11–4	11	98	11–1	98	32	6	12–11	11	99	11–8	90
33	6	11–4	15	126	14–3	126	33	6	13–1	10	77	10	76
34	5	11–5	14	121	11–8	102	34	6	13–4	14	105	14–2	106
35	5	11–5	13	110	12–8	111	35	5	13–5	14	101	10–9	80
36	5	12	13	104	10–3	85	36	7	13–5	13	100	11–1	83
37	5	12	15	113	10–11	91	37	7	13–6	14	104	13–3	98
38	6	12	15	115	13–9	114	38	4	13–7	14	101	12–6	92
39	5	12–4	12	100	11–10	96	39	6	13–8	14	110	12	88
40	7	12–4	14	109	12–3	99	40	7	13–8	15	114	12–8	93
41	7	12–5	Adult	121	15–5	124	41	7	13–8	15	110	11	80
42	4	12–6	10	85	10–9	86	42	4	14	8	50	9–3	66
43	6	12–9	11	86	12–4	97	43	4	14	9	70	9–9	70
44	5	12–10	14	106	11–9	92	44	6	14	14	101	10–1	72
45	6	12–11	15	113	12–4	95	45	7	14	15	109	12–2	87
46	6	13	11	90	11–2	86	46	5	14–1	14	99	13–7	96
47	4	13–3	11	90	10–2	77	47	6	14–1	14	100	12–3	87
48	7	13–8	15	113	11–10	87	48	6	14–1	13	98	11–9	83
49	6	13–9	14	109	11–7	84	49	7	14–3	15	113	13–9	96

TABLE 4.—*Continued*

No.	Boys						No.	Girls					
	G.	C. Age	Pt. Sc.	I. A.	T. Age	I. Q.		G.	C. Age	Pt. Sc.	I. A.	T. Age	I. Q
50	6	13–9	15	115	13–8	99	50	5	14–4	11	84	9–10	6
51	6	13–10	14	102	13–2	95	51	6	14–5	11	92	11–2	7
52	6	14	15	109	13–6	96	52	7	14–5	13	95	12–9	8
53	7	14	13	97	13–11	99	53	7	14–11	14	97	11–10	7
54	7	14	Adult	117	12–10	92	54	6	15	14	96	12–6	8
55	6	14–1	14	101	13	92	55	6	15	12	87	11–3	7
56	5	14–2	16	116	14	99	56	7	15–4	Adult	109	14–11	9
57	6	14–2	14	105	12–11	91	57	7	15–4	Adult	103	13–9	9
58	6	14–3	16	117	11–10	83	58	7	15–6	11	83	11–10	7
59	5	14–4	14	102	12–6	87	59	7	16	15	93	12–6	7
60	7	14–5	15	105	12–7	87	60	7	16–1	12	81	9–10	6
61	7	14–9	12	89	12–9	86	61	7	16–6	10	77	10–9	6
62	7	14–10	x	x	12–2	82							
63	7	14–11	15	109	13–2	88							
64	7	14–11	16	113	13	87							
65	5	15	15	99	10–8	71							
66	7	15–3	15	99	13–10	91							
67	7	15–4	x	x	14–3	93							
68	7	15–6	14	92	12–10	83							
69	7	15–6	15	101	14–8	95							
70	6	16–1	15	91	13–10	86							

* In the accompanying graphs *retarded* means that the mental age is a year or more below the chronological age; *normal,* that the mental age and chronological age are less than a year apart; *accelerated,* that the mental age is a year or more above the chronological age.

It will be seen in the case of both boys and girls that the two scales agree rather closely in designating *normal* children, but that they are far apart in the matter of *retarded* and *accelerated* pupils, especially in the upper grades. The Point Scale, for example, has only 2 boys and 4 girls retarded in grade 7, while the Terman Scale has 10 boys and 11 girls retarded in that grade. Moreover, the Point Scale has 6 boys and 6 girls accelerated in grade 7, while the Terman Scale shows 1 boy and no girls accelerated.

By general consent grade 7 was a very slow class, containing many over-aged, backward pupils. The Terman results approximate closely the actual conditions of mentality. This would indicate that the Point Scale is too easy for the upper grades. W. R. F.

[1] The estimated number of months for the Point Scale age was taken into consideration when finding the I. A.

IV

APPLICATION OF THE COURTIS STANDARD RESEARCH TESTS IN ARITHMETIC—SERIES B

Alice K. Bielaski and George Lloyd Palmer

The Courtis Standard Research Tests in Arithmetic are intended to measure ability in the four fundamental operations with integers. Too frequent use of them is not recommended as they are "neither examinations nor teaching devices." To measure the progress made in a school year under any system of instruction, it is well to use the tests at the opening of the session, at the mid-year and at the close of the year. Four forms of this series may be obtained and a different form should be used for each of the three trials. The forms are of equal difficulty and therefore the choice between them is an arbitrary one.

There is one set of tests for all grades because it is believed that true mental progress is best revealed by increased facility in the use of the same material, just as physical growth is shown by the changes in the results obtained by the use of the same measuring scales.

Test one consists of twenty-four addition examples, each containing nine addends of three digits; test two consists of twenty-four subtraction examples containing numbers of eight or nine digits; test three consists of twenty multiplication examples, each multiplicand containing four digits and each multiplier, two or three; test four consists of twenty-four division examples, each dividend containing four or five digits and each divisor, two. The digits are arranged in such a way that all the fundamental combinations are represented. Eight minutes are allowed for addition, four for subtraction, six for multiplication, and eight for division.

Full instructions for giving, scoring, and tabulating are

found in the folders accompanying the tests. The record sheet furnishes a convenient device for obtaining the class standing in terms of median speed and accuracy, as well as

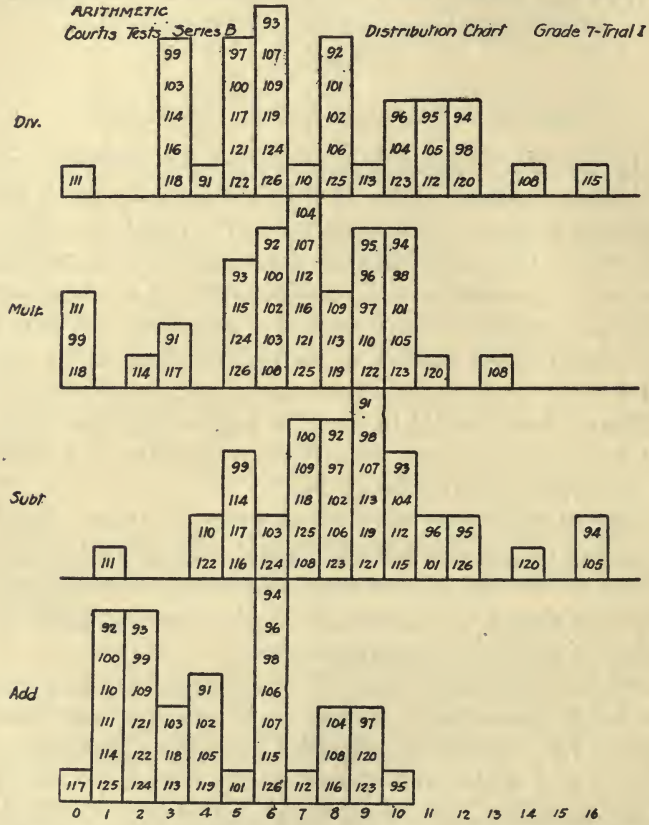


CHART XII.—Distribution of Scores, Courtis Arithmetic Tests, Series B

the percentage of efficiency. By comparison with the standard median scores in speed and accuracy, the proficiency of a grade may be determined.

The results of the tests as given in the Johns Hopkins Demonstration School are presented in Table 5.

In the first trial the 4th was the only grade proficient in

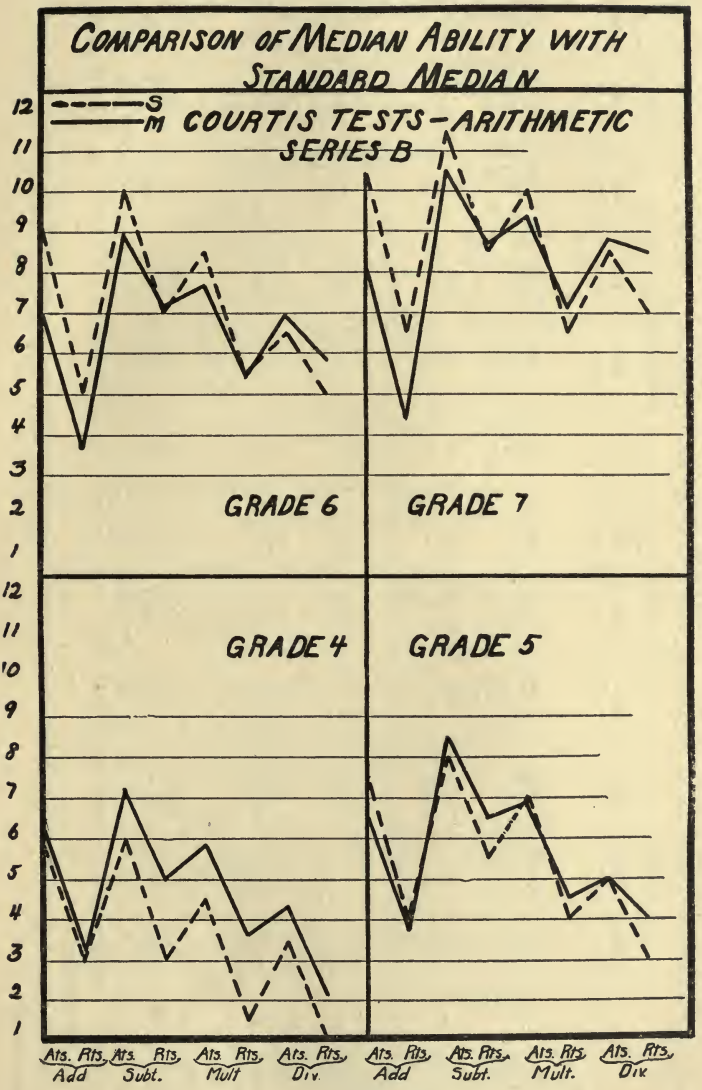

COMPARISON OF MEDIAN ABILITY WITH STANDARD MEDIAN
S
M
COURTIS TESTS – ARITHMETIC SERIES B
GRADE 6
GRADE 7
GRADE 4
GRADE 5
Ats. Rts. Add
Ats. Rts. Subt.
Ats. Rts. Mult
Ats. Rts. Div.
Ats. Rts. Add
Ats. Rts. Subt.
Ats. Rts. Mult.
Ats. Rts. Div.

CHART XIII

TABLE 5

	Addition			Subtraction			Multiplication			Division		
	Ats.*	Rts.	Eff.†	Ats.	Rts.	Eff.	Ats.	Rts.	Eff.	Ats.	Rts.	Eff.
First Trial:												
4th grade.....												
1st trial.....	6.1	3.25	10%	7.2	5.04	7%	5.8	3.6	3%	4.3	2.15	7%
Standard	6.	3.	6–6	6.	3.	7–7	4.5	1.5	6–6	3.5	1.	4–4
Deviation....	+ .1	+ .25		+1.2	+2.04		+1.3	+2.1		+ .8	+1.15	
5th grade.....												
1st trial.....	6.6	3.7	0%	8.5	6.46	4%	6.8	4.5	0%	5.	4.	11%
Standard	7.5	4.	8–8	8.	5.5	9–9	7.	4.	8–8	5	3	6–6
Deviation....	− .9	− .3		+ .5	+ .96		− .2	+ .5		0	+1	
6th grade.....												
1st trial.....	6.9	3.68	0%	9	7.2	3%	7.7	5.37	0%	6.9	5.78	9%
Standard	9	5	10–10	10	7	11–11	8.5	5.5	9–9	6.5	5	8–8
Deviation....	−2.1	−1.32		−1	+ .2		− .8	+ .13		+ .4	+ .78	
7th grade.....												
1st trial.....	8	4.35	0%	10.5	8.7	3%	9.4	7.14	5.56%	8.8	8.51	25%
Standard	10.5	6.5	11–11	11.5	8.5	12–12	10	6.5	10–10	8.5	7	10–10
Deviation....	−2.5	−2.15		−1.0	+ .2		− .6	+ .64		+ .3	+1.51	
Second Trial:												
4th grade.....												
1st trial.....	6.1	3.25	10%	7.2	5.04	7%	5.8	3.6	3%	4.3	2.15	7%
2d trial.....	6.5	3.06	0	7.4	4.94	7	5	2.7	0	4.2	2.18	7
Change.....	+ .4	− .19	−10	+ .2	− .10	0	− .8	− .9	−3	− .1	+ .03	0

* Ats. means number of examples attempted; Rts. means number of examples right. The scores in both of these columns are the medians.

† Eff. means per cent of efficiency. The standard score appears in this column under the per cent of efficiency in First Trials only.

	Addition			Subtraction			Multiplication			Division		
	Ats.	Rts.	Eff.	Ats.	Rts.	Eff.	Ats.	Rts.	Eff.	Ats.	Rts.	Eff.
5th grade.....												
1st trial.....	6.6	3.7	0%	8.5	6.46	4%	6.8	4.5	0%	5	4	11%
2d trial.....	7.3	4.1	4	13	10.4	4	6.25	3.63	0	5.5	3.7	0
Change.....	+ .7	+ .4	+4	+4.5	+3.94	0	− .55	− .87	0	+ .5	− .3	
6th grade.....												
1st trial.....	6.9	3.68	0%	9	7.2	3%	7.7	5.37	0%	6.9	5.78	9%
2d trial.....	6.6	3.47	3.22	8.66	6.67	3.22	6.9	4.55	0	7.3	6.57	22
Change.....	− .3	− .21	+3.22	− .34	− .53	+ .22	− .8	− .82	0	+ .4	+ .79	+13
7th grade.....												
1st trial.....	8	4.35	0%	10.5	8.7	3%	9.4	7.14	5.56%	8.8	8.51	25%
2d trial.....	9.3	6.88	3	12.3	10.21	0	10	8	6	9.5	9.74	9
Change.....	+1.3	+2.53	+3	+1.8	+1.51	−3	+ .6	+ .86	+.44	+ .7	+1.23	−16

all of the four fundamental operations. All the other grades were noticeably deficient in addition, the 6th grade showing practically no advance in ability over the 5th grade. However, all the grades were proficient in division. The lowest general standing was that of the 6th grade in which five of the eight deviations were negative.

In the 4th, 5th and 6th grades, the results of the second trial did not show the expected advance in ability, but this may be accounted for partially by the fact that the tests were given on a very hot and sultry day. In the 7th grade where the second trial was given on a cooler day and under more favorable conditions, there was a marked improvement in all four operations. It is believed that under similar circumstances, the tests given in the other grades would have shown satisfactory advancement.

On the whole, it would seem that subtraction and division are better learned than addition and multiplication and that these two latter processes should receive more emphasis in teaching. But it may be well to consider whether the addition and multiplication tests are well standardized. Would examples containing fewer addends furnish a better scale of measurement in addition?

The Courtis definition of efficiency is somewhat misleading. To be efficient a pupil must be 100 per cent accurate and maintain a speed equal to, or greater than, the standard speed. This gives no credit to one who has exceeded the standard in attempts and rights, but has fallen below 100 per cent in accuracy. *E.g.,* in test No. 2, Subtraction, there was one pupil in the 5th grade who correctly solved twenty-two out of twenty-four examples. The standard score in this test is 9—9, and if the pupil had attempted 9 examples and had solved them correctly, he would have been efficient, but having attempted twenty-four and solved but twenty-two correctly, he falls below efficiency. Is this grading fair?

These tests are valuable tools for making rough measurements. Tests consisting of carefully graded examples are finer instruments for diagnosing individual ills and hence furnish a better guide to the teacher in selecting a remedy.

V

RESULTS IN ARITHMETIC BY WOODY SCALE "A"

W. H. Davis and R. L. Clark

The Woody Arithmetic Scale, Series A, developed in 1915–16 by Clifford Woody from about 20,000 test sheets of pupils in seven different school systems in Indiana, New Jersey, Connecticut, and New York, aims to test pupils, classes, schools, and systems in accuracy in the four fundamental operations.

The four tests contain 148 problems. Those of each test are of increasing difficulty and each problem is of fixed value. The progress from grade to grade, therefore, can be definitely determined. Twenty minutes is allowed for each test, and absolute accuracy is the basis for scoring. The judgment of the scorers is thus eliminated.

The scale was used by the writers in testing the fourth, fifth, sixth, and seventh grades in the Demonstration School of Johns Hopkins University in July, 1917. The pupils in this school, in general, failed of promotion during the preceding year and were attending the school to prepare for the next grade. This condition and the fact that Baltimore allots an unusual amount of time to Arithmetic would not warrant the expectation of class medians and scores higher than the standards derived by Woody for grades in the middle of the school year.

From the tabulations of the original data, showing the particular and total problems solved by each pupil, and the number of pupils who solved each problem, were obtained the following distribution tables, medians, and scores.

COMPARISON OF CLASS AND STANDARD MEDIANS

	Addition			Subtraction		
	Class	Standard	Deviation	Class	Standard	Deviation
Grade IV	20.9	18.3	+2.6	19.7	15.7	+4
Grade V	28.5	23	+5.5	24.7	20.4	+4.3
Grade VI	29.5	29.7	— .2	28.6	24.9	+3.7
Grade VII	32	32.4	— .4	30.6	28.5	+2.1

	Multiplication			Division		
	Class	Standard	Deviation	Class	Standard	Deviation
Grade IV	17.5	11	+6.5	15.5	9.9	+5.6
Grade V	23	18.3	+4.7	22.5	16.5	+6
Grade VI	29.3	26.1	+3.2	25.9	23.8	+2.1
Grade VII	32.5	30.6	+1.9	26.9	27.4	— .5

COMPARISON OF CLASS AND STANDARD SCORES

	Addition			Subtraction		
	Class	Standard	Deviation	Class	Standard	Deviation
Grade IV	6.76	6.11	+ .65	4.95	4.22	+ .73
Grade V	7.77	6.99	+ .78	6.82	5.47	+1.35
Grade VI	8.20	7.95	+ .25	7.04	6.46	+ .58
Grade VII	8.17	8.65	— .48	8.01	7.31	+ .70

	Multiplication			Division		
	Class	Standard	Deviation	Class	Standard	Deviation
Grade IV	5.44	4.05	+1.39	4.65	3.21	+1.44
Grade V	6.10	5.53	+ .57	5.42	4.94	+ .48
Grade VI	7.28	6.72	+ .56	6.11	5.87	+ .24
Grade VII	7.16	7.26	— .10	6.51	6.59	— .08

The class score is the value of that problem which 50 per cent of the class can solve correctly and is obtained by getting the average value of the five problems which came nearest to being solved by 50 per cent of the class. Thus in subtraction in grade IV:

Problem	Per Cent. Solving	Est. Value	Correction		Value
15	54	3.70	+ .15	=	3.85
17	65	4.41	+ .57	=	4.98
18	69	4.42	+ .74	=	5.16
19	38	5.18	— .45	=	4.73
22	57	5.57	+ .26	=	6.01
					5)24.73
				Class Score	4.95

CONCLUSIONS

1. The distribution sheets disclose overlapping of grades as indicated in the following table:

PERCENTAGE OF OVERLAP IN NUMBER OF PROBLEMS SOLVED

	Addition		Subtraction		Multiplication		Division	
	Above Median of Next Higher, Per Cent.	Below Median of Next Lower, Per Cent.	Above, Per Cent.	Below, Per Cent.	Above, Per Cent.	Below, Per Cent.	Above, Per Cent.	Below, Per Cent.
Grade IV	0		0		10		6.6	
Grade V	48	14	15.4	3.8	18	10.4	28.6	7
Grade VI	38	35	16.6	6.6	29	3	42	13
Grade VII		21		22.5		22		38

2. Comparisons of class with standard medians and scores show that the fourth grade far excels the standards and that these deviations, decreasing, in general, in the succeeding grades, become negative deviations in two medians and three scores of the seventh. The facts that in Baltimore arithmetic is taught in the first grade and that more time is allotted to arithmetic in the first four grades than in any city other than Cincinnati would account for the excellence of the fourth grade, and the greater retardation in the upper grades would account in great part for the lower achievements of the seventh.

3. The practical use of the scale is conditioned by the amount of time required to tabulate the results and calculate the final scores.

VI

AN EXPERIMENT IN MEASURING THE HAND-WRITING OF A GROUP OF CHILDREN FOR SPEED AND QUALITY

WILLIAM R. FLOWERS

One hundred eighteen children were included in this survey of hand-writing measured for speed and quality, comprising grades 4, 5, 6, and 7.

Adopting the plan of Starch in his "Educational Measurements," each pupil was given exactly two minutes in which to write as often as possible the sentence "Mary had a little lamb," in order to find the number of letters written per minute. Immediately following this test, each pupil wrote a paragraph from dictation, the vocabulary of which was simple enough for even the youngest. This was the test for quality. The writing was done on lined paper of uniform size, with pen and ink. Nothing was said to the pupils to urge unusual speed or unusual carefulness.

In measuring the quality three scales were used; viz. the Ayres, the Thorndike, and the Freeman. Each specimen was graded first by the Ayres scale and the number recorded so that the examiner could not see it, when, twenty-four hours later, the same specimen was graded by the Thorndike scale. After another interval of twenty-four hours each specimen was again graded by the Freeman scale, five points of judgment being recorded for the last scale—uniformity of slant, uniformity of alignment, quality of line, letter formation, and spacing. The Ayres scale grades the quality on a percentage basis from twenty to ninety. The Thorndike scale grades the quality by a series of numbers from 4 to 18. The Freeman scale grades the quality by giving a number from 1 to 5 for each of the above

mentioned five characteristics, the total of these marks, 5 to 25, being the final rating. For convenience of comparison with the Ayres scale each of the others has been changed to a percentage basis.

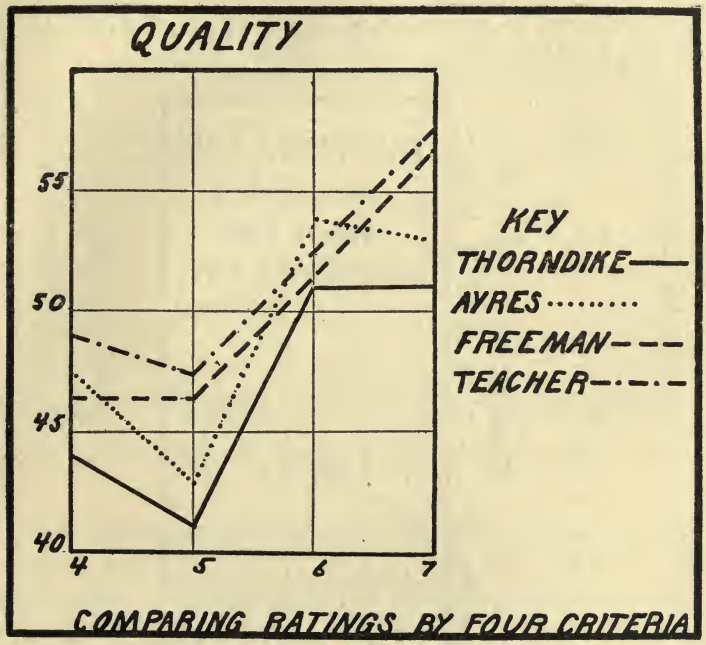


CHART XIV

After the three measuring scales had been used, the grade teacher was asked to put her estimate on her pupil's papers, without, however, using any measuring scale.

The following summary gives for each grade the average rating for quality by each of the preceding four criteria:

Grades	Ayres Scale	Thorndike	Freeman	Teacher's Estimate
4	47.6	44.0	46.4	49.0
5	43.9	41.0	46.1	47.5
6	53.9	51.0	50.9	52.5
7	53.0	51.0	56.3	56.7

It will be seen from the above ratings that the Thorndike scale gives the lowest rating in each grade except the sixth; that the average of the Ayres and Freeman ratings is almost the same (49.6 for Ayres, 50.2 for Freeman); and that, except in grade 6, the teacher's estimate is higher than that obtained by applying any scale. In graphic form these results are shown thus:

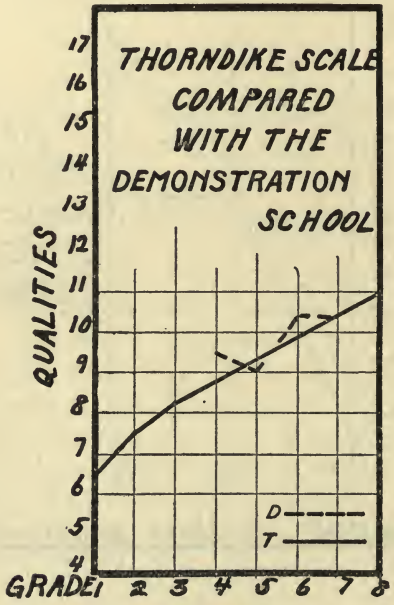


CHART XV

The accompanying graphs (charts XV and XVI) compare the average quality of the classes by the Thorndike scale with the standard curve for that scale as given by Starch. It will be noticed that only grade five falls below, and that grades four and six are above.

In the summary it will be noted that there was not a very wide difference in the average quality of the writing of any grade as compared with the grade above or below it. This is a usual result, and is shown more strikingly by the following distribution charts.

The Freeman scale was chosen as the basis of this comparison, because the experimenter thought that the detailed criticism of quality of writing afforded by that scale would be the fairest criterion. It is probable that the same overlapping would occur with any other scale.

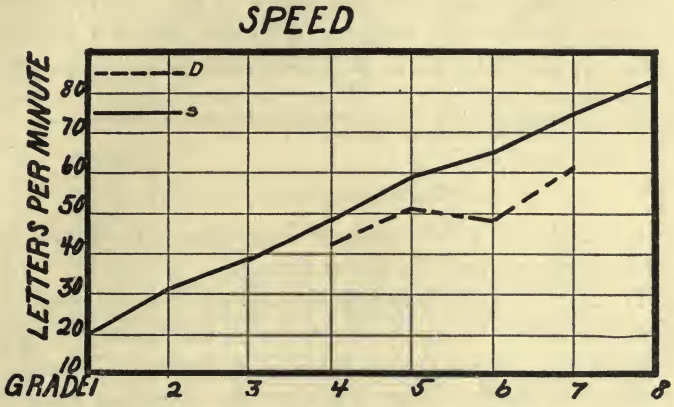


CHART XVI.—Starch's Standard Compared with the Demonstration School

Overlapping in speed also occurs, there being, for example, 7 pupils in grade 4, 12 in grade 5, 10 in grade 6, and 11 in grade 7 all writing at the rate of 45–54 letters per minute. This is 34 per cent of the whole number tested.

Correlations

The inquiry naturally presents itself as to the correlation between quality and speed. As might be expected there is a negative correlation of —.38, with probable error of .07; i. e., the higher the quality the lower the speed.

Another interesting question is the correlation between the teacher's estimate and the experimenter's. In grade 6 the teacher used four of Freeman's five points of judgment (without a knowledge of Freeman's scale), and the correlation for that scale between her estimate and the experimenter's (by the Spearman formula) was +.57. The

teacher of grade four was familiar with the Ayres scale, and the correlation between her estimate and the experimenter's was + .62.

The teachers of grades 5 and 7 had never used any measuring scale. In grade 5 the correlation between the teacher's estimate and the experimenter's by the Ayres scale was + .47; between the teacher's estimate and the Freeman scale, + .50. In grade 7 the correlation between the teacher's estimate and Ayres scale and between teacher's estimate and Freeman scale was the same, + .47.

Criticism and Conclusions

I. Quality of handwriting is very difficult to measure accurately by any scale in use at present.

II. The Ayres scale is difficult to apply because it does not contain sufficient variety of specimens, and because it passes a composite, rather than a detailed, judgment. It essays to measure *legibility*. Even if it does this (which is doubtful) few teachers are satisfied with writing that is merely legible.

III. The Thorndike scale is superior to the Ayres in that it contains more specimens. On the other hand, it seems to the writer unfortunate that the specimens are not distributed more equally among the various qualities (quality 10, e. g., having but one illustration), and that the system of grading is so inconvenient. One wishes that Professor Thorndike would arrange a new scale with the above difficulties eliminated.

IV. The Freeman scale seems the most rational because it itemizes the characteristics of good and bad writing and judges each separately. Its judgment is detailed and specific, not composite. It is also most practical in pointing out to pupils exactly the faults in their writing. Value would be added to this scale if five grades of quality instead of three were given and more specimens in each grade included.

V. As shown by the coefficients of the correlation in grades 4 and 6 on the one hand and grades 5 and 7 on the

other, *some* measuring scale is better than *none* in assisting to a uniform grading of quality of writing.

VI. In view of the amount of overlapping in quality as shown in the distribution charts, it seems desirable to have *grade* measuring scales instead of one scale for all grades. Under such a plan a pupil in any grade who reaches the maximum of quality for his grade could be excused from further formal drill in writing, unless his writing dete-

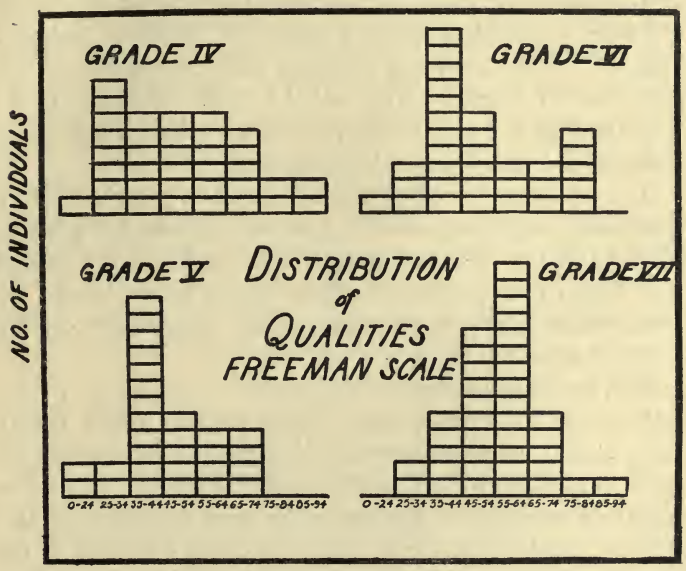


CHART XVII

riorated. This would recognize individual differences and enable those with capacity for good writing to stress some other subject in which they might be deficient. This plan has been followed for the past two years with good results in the school with which the writer is connected.

VII. Since it is not enough merely to write well, reasonable speed also being demanded, it is desirable in all scales to combine quality and speed and give one rating to include *both*. This has not yet been satisfactorily worked out.

VII

THE KANSAS SILENT READING TEST

MARY O. EBAUGH

The Kansas Silent Reading Test is designed to measure the ability "to interpret the meaning of sentences and paragraphs." The two factors—speed and comprehension—are combined in a single mark, and the child's ability to read is measured by the number of reading exercises which he can comprehend accurately within a given time.

The test includes three sets of exercises—one for grades 3, 4 and 5; one for grades 6, 7 and 8; and one for grades 9, 10, 11 and 12. The first exercise for grade 3 is as follows:

"I have red, green and yellow papers in my hand. If I place the red and green papers on the chair which color do I still have in my hand?"

The last exercise in the last set is:

"At sea level water boils at 212 degrees above zero on the Fahrenheit thermometer, and at 100 degrees above zero on the Centigrade thermometer. The zero point on the Centigrade thermometer represents the same temperature as 32 degrees on the Fahrenheit thermometer. A change in temperature which would raise the mercury in a Centigrade thermometer 5 degrees would raise the mercury in a Fahrenheit thermometer how many degrees?"

Each exercise contains not less than 15 words; few contain more than 60. Each is supposed to be subject to only one interpretation and to call for but one thing so that what the child does in response to it will be wholly right or wholly wrong. Each is so planned as to reduce written interpretation to a minimum so as not to confuse ability to get meaning with ability to reproduce meaning.

The value of each exercise indicates the relative length of

time required on the average by children of a certain grade to answer the exercise correctly.

Some of the exercises are short and easy to remember; others are more difficult. Some require direct reasoning, while others are of the nature of a puzzle.

The answers indicate in many cases that the pupils fail to answer correctly although apparently they comprehend the statement. Since the test is not designed as a memory test, a test in reasoning, or a test in solving problems, but a test in which "the difficulty of each exercise must depend upon the child's interpreting the English language," the difficulties connected with memorizing, reasoning or the solution of problems should be kept as far as possible on an equal plane and difficulties in vocabulary or in construction should be the basis upon which the increased difficulty of one exercise over another should depend.

Revision of a few of the exercises, which are not stated clearly, would add value to the scale. It is wrong to rank answers indicating partial comprehension in the same way as those indicating no comprehension at all. Furthermore, the test would be much more valuable if it had been planned to reveal specific causes of strength or weakness in each individual effort. It is impossible to tell whether low scores are due to slowness or to lack of comprehension.

The test is definite and practical; it takes a short time to give and can be given to large numbers at the same time. In spite of its limitations it furnishes instructive data.

The results of the tests given in the Demonstration School closely approximate the results established by Kelly in an examination of 9,252 children in 19 cities of Kansas. The median for the 4th grade was .2 higher than the standard; the median for the 5th grade was .6 higher than the standard; for the 6th grade it was .4 lower and for the 7th 1.7 lower.

There was great variability in ability among the boys, among the girls, and in the classes as a whole. In each class the average and median scores for the boys were higher than those for the girls.

In the 4th grade 40 per cent of the boys and 60 per cent of the girls were below the class median; in the 5th grade 47.3 per cent of the boys and 55.5 per cent of the girls were below the class

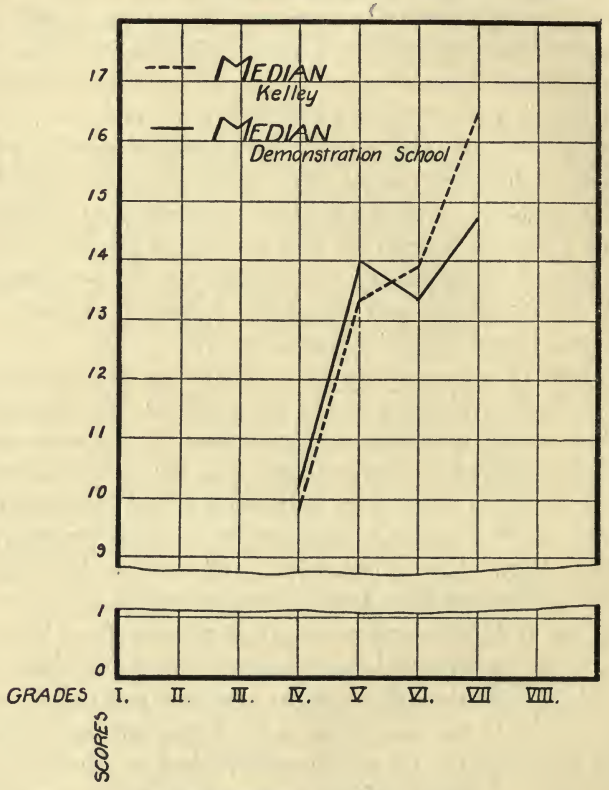


CHART XVIII.—Silent Reading

median; in the 7th grade 42.1 per cent of the boys and 55.5 per cent of the girls were below the class median.

The overlapping of grades was very noticeable particularly in the 6th and 7th grades. More than 45 per cent of the 6th grade made a score higher than the median score of the 7th grade. More than 40 per cent of the 7th grade made a score

lower than the median score of the 6th grade. 20 per cent of the 4th grade made a score higher than the median score of the 5th grade, and 14 per cent of the 5th grade made a score lower than the median score of the 4th grade. The 5th and 6th grades could not be compared because different tests were given in these two grades.

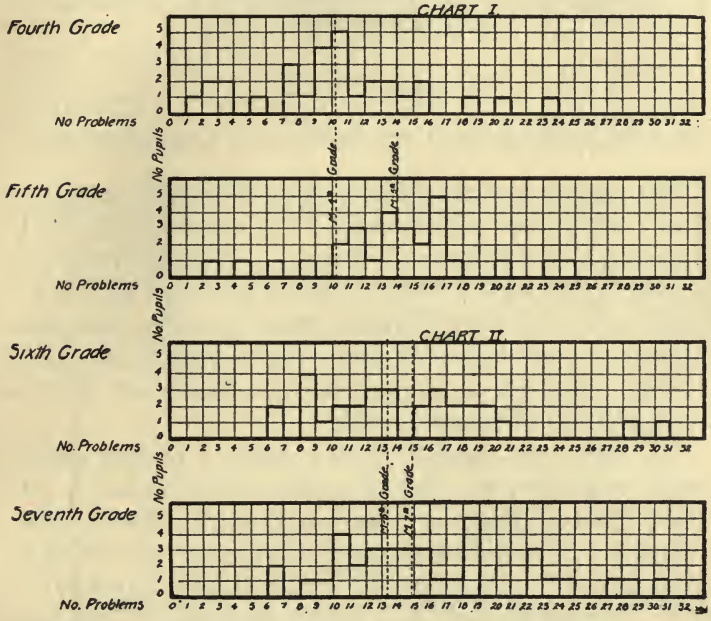


CHART XIX.—SILENT READING. CHARTS SHOWING OVERLAPPING OF GRADES

The value of the scores made did not vary directly with age. In the test given to the 6th and 7th grades the highest score was made by the youngest child and after passing the normal age for children in these grades the average score rapidly became lower. In the test given to the 4th and 5th grades the average scores did not show such a steady decrease for pupils above the normal age, but showed great variability though the average score for the highest age was much lower than that for any other age.

VIII

THE STARCH TEST FOR SPEED AND COMPREHENSION AND THE THORNDIKE VISUAL VOCABULARY TEST

BYRON J. GRIMES

A. THE STARCH TEST FOR SPEED AND COMPREHENSION

The reading test, Series A, used in this survey measures speed and comprehension only. The speed of reading is measured by the number of words of a certain text that can be read in one second. The ability to reproduce text is accepted as a measure of comprehension. To state this more clearly: the number of words written immediately after reading, containing or reproducing the thought of the text is the measure of comprehension.

The entire test consists of nine pages of reading matter suited to the first nine school grades and advancing in difficulty from one selection to another by fairly uniform steps.

Pupils are instructed to read for just 30 seconds with as much speed as will permit of their understanding what is read. It must be made clear before beginning to read that they will be required to turn the sheet over and write on the back as much of the story as they can recall. Each pupil must begin and stop his reading at exactly the same time. No time limit is set for reproduction.

Illustration: "Once upon a time there was a rich man and a king who had a daughter named Midas."

It has not yet been clearly demonstrated that the ability to reproduce in writing what has been read is a fair estimate of reading ability. Oral reproduction, while offering some difficulties for the teacher, would simplify the process for the children tested.

The small number of words read by most children would seem to indicate that a longer period than 30 seconds would be preferable.

A summary of the results of the Starch Reading test, in median scores, is as follows:

	Speed	Deviation from Standard	Comprehension	Deviation from Standard	No. Words Read	Total Score
Fourth grade	2.1	− .3	29	+ 1	64	27
Fifth grade	2.5	− .3	28	− 5	81	32
Sixth grade	2.6	− .6	24	−14	86	28
Seventh grade	2.1	−1.5	32	−13	62	29

It is evident from the scores shown above that the children of this school are below standard in rate of reading and also in comprehension.

This is further evidence of the generally accepted theory that slow reading and poor comprehension are closely related.

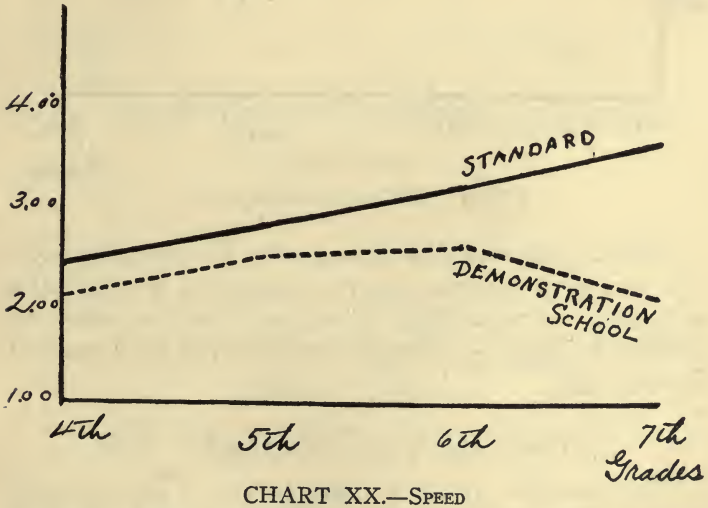


CHART XX.—SPEED

A comparison of graphs 1 and 2 makes an interesting study relative to the seventh grade. In speed this grade shows a decided falling off, but in comprehension a rela-

tively good score. (This may be due to larger experience and wider reading.) This ability to interpret may be due to the enlarged experience and wider range of reading of two and three, even four, additional years in school.

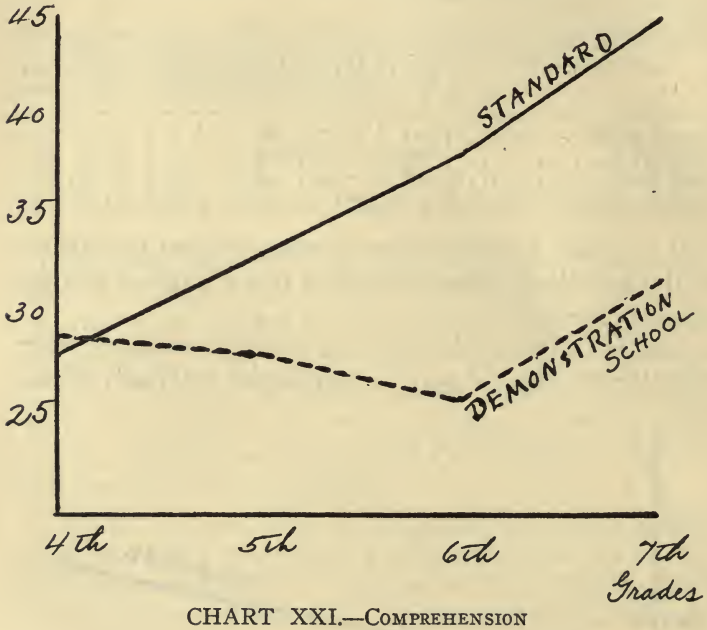


CHART XXI.—Comprehension

Difficulty was experienced in scoring for comprehension. The element of judgment enters so largely as to permit of a wide range of variability. Ideas got from text read but changed in order or arrangement could not be considered reproduction.

B. Thorndike Visual Vocabulary Test

The Thorndike Visual Vocabulary Test is an attempt to measure silent reading so far as it concerns the understanding of words singly, unconfused with the ability to express one's self orally or in writing.

The test consists of 9 lines of 5 words each, with the ex-

ception of the last line which contains only 3 words. All words on the same line are supposed to be equally hard to understand and the difficulty increases in equal amounts from line to line except that the difficulty between the 8th

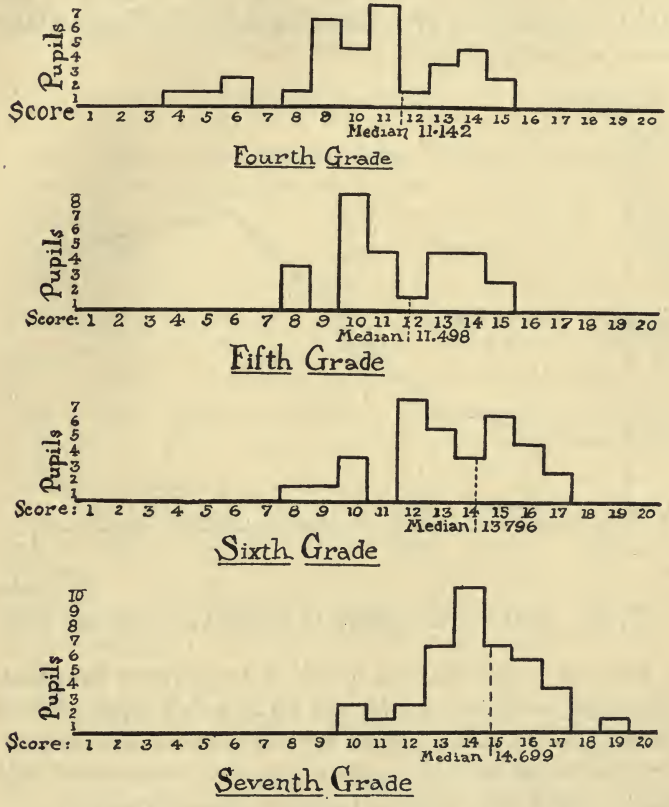


CHART XXIa.—Distribution of Scores for the Trabue Completion Test.

line and the line preceding and following is only half as great as the difference between any two succeeding lines. The test consists in the correct listing of each of the words in the 9 lines according to a definite classification laid down in the direction for taking the test. The time element is not considered at all.

The scale really measures the ability to understand printed words only well enough to classify them and it does not test a pupil's knowledge of a word in its natural setting. Other limitations of the scale are: (1) undue predominance is given to names of animals and flowers, (2) the omission of pronouns, conjunctions, prepositions, auxiliary verbs and other words expressing relation.

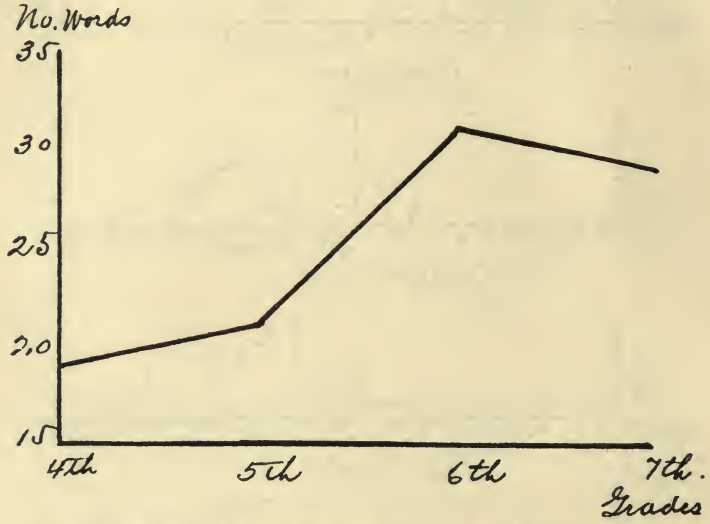


CHART XXII.—ACHIEVEMENT AS SHOWN BY THORNDIKE TEST

There is a close relation in the findings of all three tests. The fourth and fifth grades are making fair progress while the sixth and seventh grades are much below standard.

Starch's test for speed shows that the entire school reads slowly, which may account for poor comprehension.

That the entire school is made up largely of retarded pupils is a possible explanation of the low scores obtained. This can be fairly well determined by a study of the tests in all school subjects.

Attention is called to the decided overlapping in grades. Of the 31 pupils in the sixth grade 14 could just as well be placed in the seventh; six of the fourth grade could do the work in reading of the fifth grade.

IX

APPLICATION OF AYRES, BUCKINGHAM, AND STARCH SCALES IN SPELLING

DOROTHY B. BERRY

The Ayres Scale is based upon the one thousand most common words in the English language. These words were selected by combining the results of four previous investigations, which had as their object the selection of the words most commonly used in different sorts of writings. The first study was made by the Rev. J. Knowles in 1904 in a pamphlet entitled, "The London Point System of Reading for the Blind." From passages in the English Bible and from various authors, containing 100,000 words, a list was made of the 353 words which occurred most frequently. The second study was made by R. C. Eldridge and the results were published in 1911 in "Six Thousand Common English Words." The frequency of different words was made on a basis of an analysis of 250 articles taken from issues of four Sunday newspapers in Buffalo. These articles counting repetitions contained 43,989 words.

The third study was made by L. P. Ayres in 1913 and results were published in "The Spelling Vocabularies of Personal and Business Letters." The study consisted of the tabulation of 200,000 words taken from family correspondence of 13 adults. The total vocabulary consisted of 5,200 different words. The list of one thousand words finally selected was determined by finding the frequency with which each word appeared in the tabulation of each study, weighting that frequency to the size of the base, adding the four frequencies and finding their average.

The 1,000 words were first made up into 50 lists of 20 words each and these lists were then given to various grades

in the schools of 84 cities. The data secured from these tests made an aggregate of 1,400,000 spellings by 70,000 children. The results constitute the basis of the scale.

The scale explains itself. It is divided into twenty-six columns lettered from A to Z. All the words in each are of approximately equal spelling difficulty. The steps in spelling difficulty from each column to the next are approximately equal steps. The numbers at the top of the scale indicate about what per cent of correct spellings may be expected among the children of the different grades. By means of these groupings a child's spelling may be located in terms of grades.

The Starch Tests were selected in the following manner: The first defined word on every even-numbered page in Webster's New International Dictionary was chosen, making a total of 1186 words. From these all technical, scientific, and obsolete words were discarded, leaving 600 words. These were then arranged alphabetically in the order of size, beginning with three letter words down to the longest. This list was then divided into six lists of 100 words each, by choosing for the first list the 1st, 7th, 13th, etc.; for the second list the 2d, 8th, 14th, etc.; and so on until the sixth list was completed. These tests have been standardized by administering them to 2,500 pupils in 12 schools of 5 cities. The average results have been tabulated.

The Buckingham list was selected in the following manner: From a list of 5,000 words, taken from five Spelling Books a list of 270 words was used for a test. This was called the "Original List." These words had to satisfy the following requirements: (1) All of them had to be words in the speaking vocabulary of a third grade child, and (2) spelling difficulty of many of them had to be great enough to test the ability of eighth grade children. These were then placed in a continuous passage and the whole dictated to different grades. Two measurements were recorded: (1) the number of times each word was correctly spelled in each grade, and (2) the percentage of the entire number of words each pupil spelled correctly in each grade.

The basis upon which the "Selected List" was chosen is as follows: Referring to the previous study it was seen that the word *across* was spelled by 17 per cent of the third grade children, which means that it was not too hard to serve as a test of their ability. By the time the eighth grade was reached it still served as a test of ability. Thus 100 words were selected.

These words were again put into sentences and from the data collected two lists were then selected, each containing 25 words, which show a regular increase in difficulty as we pass from grade to grade. These are known as the "First Preferred List" and "Second Preferred List." In this way Buckingham has provided a basis of comparison, as a method of testing the relative ability of different classes.

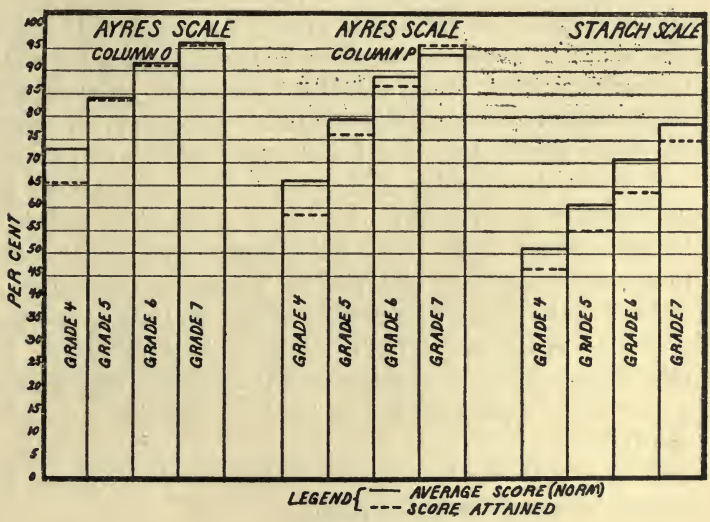


CHART XXIII

In the first Ayres' test the grades attained the average score with only one deviation of —9 in the fourth grade. In the second Ayres' test there was only a slight negative deviation in the three lower grades, but in the seventh grade, the deviation was + 2.

No word in the Ayres' tests was missed more than forty times nor less than five times. In the Starch test some words were missed 120 times or more.

The Starch average scores are rather low and none of the grades attained the average. The deviation ranged from —4 to —7.

In constructing a test for any grade only the crucial words should be used. Crucial words are believed to be those which may be spelled by 50 per cent of the pupils of that grade, or those words of approximately equal difficulty of which the average score of the pupils of that grade will be 50 per cent. Buckingham used the percentage. In this respect the Ayres and Buckingham methods are superior to the Starch method.

The Starch scale is merely a random selection of words with no regard for the child's writing vocabulary. It seems that any test would be valueless in testing words the child has never studied. Such words are *nunciature, bizarre,* and *ineffectuality* serve as good examples. In scoring, all words are of the same value. *But* has the same value as *nunciature.*

It seems that spelling ability can hardly be measured by an arbitrary list of words. No list of 50 words is sufficient to test spelling ability. The Buckingham scale might serve to test large groups of children, but hardly the individual.

Ayres has scaled a foundation spelling vocabulary and has presented groups of words of equal spelling difficulty. In this respect Ayres is superior to the others in that he has presented a representative basic list, consisting of 1,000 words.

In each test, the boys scored as high as the girls and in several instances surpassed them.

X

THE TRABUE COMPLETION TEST

MAYNARD A. CLEMENS AND FRANKLIN E. RATHBUN

As defined by the originator, the Trabue sentence completion test is an index of language ability. It may also be considered as a test of ability in logical thinking. Professor H. Ebbinghaus who devised the paragraph completion test, of which the present test is a direct lineal descendant, stated that it constituted a real test of intelligence. Other psychologists have identified it as a test of "association," "memory," and "imagination."

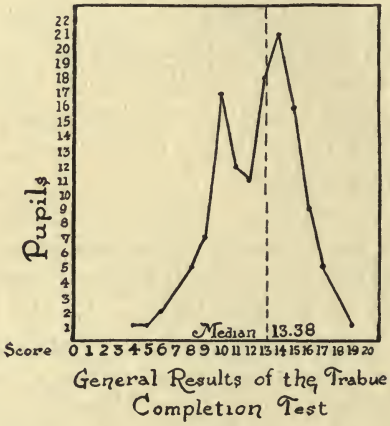
This test consists of a number of sentences having one or more blank spaces where words have been omitted. The students are called upon to write the most appropriate words they can think of in these blanks. No list of words has been arbitrarily determined upon in advance to be supplied; hence, in most cases, there is an option of several words. The sentences are of progressive difficulty, permitting only the survival of the fittest. The first few are of such a character that little difficulty will be experienced in supplying the missing word even by the lower grade students, whereas the last often baffle the wits of mature men and women.

An arbitrary system of scoring is employed. Mistakes of orthography are not considered; simply the aptness of the words filled in is judged. Considering carefully the context if the sentence has been completed satisfactorily, a score of 2 is given; if slight grammatical mistakes occur or infelicitous words have been used, a score of 1 is given; but if a wrong word has been employed, making an utterly hopeless expression, zero is assigned.

In devising this test, Dr. Trabue took fifty-six incomplete sentences of graduated difficulty and in 1914–15 secured re-

sults from several thousand students of New York and New Jersey. A careful evaluation of these sentences was made; and, as a result, many were discarded. Twenty-four were retained and graded, constituting a new test. This was called scale A, and during 1915 it was given to 6,000 students of New York, New Jersey and many middle Western states.

Since then, scale A, which was too cumbersome and required too much time for presentation, has been formulated



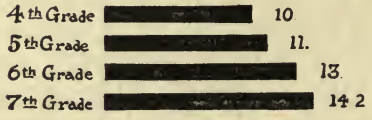


CHART XXIV

into a series of smaller scales of ten sentences each requiring only 5 to 7 minutes for testing.

We have tested with scale B one hundred twenty-six students distributed in the fourth, fifth, sixth and seventh grades of the Demonstration School.

In grading the papers the grading given by the Trabue

monograph was considered a final court of appeal. It is interesting to note that the list of words given there as answers of school children for this test conformed quite generally to the list secured by us.

Following are the general results:

DISTRIBUTION OF TOTAL SCORES FOR SENTENCES

Grade	Students	1	2	3	4	5	6	7	8	9	10	Total
ı grade	32	58	62	60	54	45	27	14	8	2	0	330
grade.....	26	43	50	50	50	45	24	18	10	7	0	297
grade.....	32	60	64	62	62	51	54	31	22	16	2	424
h grade.....	36	72	72	70	72	69	57	32	39	18	10	513
	126	233	248	242	238	212	162	96	79	43	12	1562
l average per vidual.....		1.8	1.9	1.9	1.8	1.6	1.2	.7	.6	.3	.09	12.5

.....0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
f Students																					
ining score..0	0	0	0	1	1	2	0	5	7	17	12	11	18	21	16	9	5	0	1	0	126

Median 13.38

The originator believes that this scale will mark quite definitely the intelligence for each grade. This being true we should expect a progressive increase in the average and an advancement in the position of the median from group to group. This rate of progress, too, should be fairly well fixed.

In general these conclusions are substantiated by the results secured by us.

Grade	Estimated Median	Actual Median	Difference	Average Score
Fourth.....	8.0	10.75	2.75+	10.
Fifth.....	9.6	11.5	1.9 +	11.
Sixth.....	11.0	13.75	2.75+	13.
Seventh.....	12.3	14.63	2.3 +	14.2

The better showing of these students than the calculated estimates for similar grades is probably due to the comparatively smaller number which is less affected by extremes and to the fact that these pupils are about ready for a grade higher than that in which they are now classified.

The value of the scale should be considered. Is it worth while? Does it lay bare definite language faults to enable a cure to be administered? It is doubtful if it brings results which could be obtained by other tests. Certainly as a language test, it leaves much to be desired. No provision has been worked out for determining the results qualitatively. With it, no one can exactly diagnose the student's troubles; nor is the cure very plain. To obtain a higher score does a student need more grammar, more reading, more spelling, literature, a larger vocabulary? Probably all of these, as higher scores seem to mark advanced education. It may be better suited for testing general intelligence. Hence, we must conclude that it is probably better correlated with other tests. Alone, it is simply an index.

XI

HILLEGAS SCALE FOR THE MEASUREMENT OF QUALITY IN ENGLISH COMPOSITION

J. B. H. Bowser and H. L. Rinehart

This test in composition was given to grades four, five, six, and seven of the Johns Hopkins University Summer Demonstration School.

A few minutes were allowed the pupils to place on the paper the name, date, age, grade, and school. The subject was *The Season that I Like Best and Why.* The time given for writing the composition was fifteen minutes.

The reading and evaluation of the compositions were made independently by the writers.

Table 6 shows the qualities and steps into which each judge placed each composition of the four grades. It will be seen that, in many cases, the scores given by the individual judges place the compositions in the same step. When this did not occur, the average of the score given a composition by the two judges was taken as the final score for that paper and the composition was placed in the step to which that final score belonged.

Table 7 gives the grade distribution in which overlapping of the scores in the several grades is apparent. The Standard Medians as given by Starch are: for the fourth grade, 26; for the fifth grade, 31; for the sixth grade, 36; and for the seventh grade, 41.

Table 8 shows that the median of grade four is 23, a deviation of —3 from the standard, of grade five, 30; a deviation of —1; of grade six, 36, no deviation, and of grade seven, 44, a deviation of +3. This table also shows the medians as given by the individual judges. The medians

TABLE 6

COMPARISON OF THE JUDGES' MARKS AND THE FINAL STEP INTO WHICH INDIVIDUAL COMPOSITIONS ARE PLACED

1						2						3						4					
Grade IV.						Grade V.						Grade VI.						Grade VII.					
Comp. Number	First Judge		Second Judge		Final Step	Comp. Number	First Judge		Second Judge		Final Step	Comp. Number	First Judge		Second Judge		Final Step	Comp. Number	First Judge		Second Judge		Final Step
	Quality	Step	Quality	Step			Quality	Step	Quality	Step			Quality	Step	Quality	Step			Quality	Step	Quality	Step	
1	25	2	18	1	1	1	35	3	30	2	3	1	26	2	40	3	3	1	38	3	55	5	4
2	21	1	12	1	1	2	37	3	40	3	3	2	53	5	45	4	4	2	33	3	57	5	4
3	23	2	24	2	2	3	32	3	36	3	3	3	30	2	28	2	2	3	30	3	32	3	3
4	25	2	22	2	2	4	38	3	26	2	3	4	28	2	26	2	2	4	34	3	50	4	4
5	28	2	25	2	2	5	30	2	31	2	2	5	60	5	50	4	5	5	56	5	60	5	5
6	37	3	50	4	4	6	28	2	25	2	2	6	35	3	32	3	3	6	36	3	38	3	3
7	23	2	20	1	1	7	20	1	20	1	1	7	58	5	50	4	5	7	23	2	40	3	3
8	28	2	18	1	2	8	34	3	40	3	3	8	30	3	40	3	3	8	29	2	38	3	3
9	26	2	30	2	2	9	35	3	28	2	3	9	55	5	50	4	4	9	45	4	40	3	4
10	24	2	28	2	2	10	20	1	22	1	1	10	18	1	27	2	2	10	40	3	40	3	3
11	28	2	30	2	2	11	21	1	25	2	2	11	34	3	40	3	3	11	33	3	40	3	3
12	23	2	18	1	1	12	23	2	20	1	1	12	37	3	45	4	3	12	40	3	40	3	3
13	26	2	30	2	2	13	23	2	20	1	1	13	28	2	33	3	2	13	40	3	49	4	4
14	27	2	28	2	2	14	18	1	15	1	1	14	37	3	30	2	3	14	37	3	32	3	3
15	24	2	20	1	2	15	25	2	26	2	2	15	34	3	32	3	3	15	45	4	48	4	4
16	25	2	25	2	2	16	45	4	43	4	4	16	35	3	47	4	3	16	37	3	35	3	3
17	25	2	25	2	2	17	21	1	25	2	2	17	20	1	28	2	2	17	60	5	54	5	5
18	23	2	0	0	1	18	24	2	26	2	2	18	25	2	26	2	2	18	55	5	58	5	5
19	22	2	18	1	1	19	37	3	35	3	3	19	59	5	30	2	4	19	32	3	40	3	3
20	24	2	25	2	2	20	40	3	50	4	4	20	30	2	30	2	2	20	30	3	38	3	3
21	23	2	20	1	1	21	33	3	35	3	3	21	36	3	42	3	3	21	38	3	55	5	4
22	24	2	18	1	1	22	25	2	30	2	2	22	35	3	48	4	3	22	39	3	38	3	3
23	25	2	25	2	2	23	28	2	30	2	2	23	32	3	38	3	3	23	36	3	40	3	3
24	25	2	20	1	2	24	23	2	25	2	2	24	48	4	40	3	4	24	54	5	55	5	5
25	23	2	20	1	1	25	30	2	49	4	3	25	28	2	35	3	3	25	32	3	48	4	3
26	25	2	18	1	1	26	20	1	31	2	2	26	37	3	48	4	4	26	73	7	47	4	5
27	23	2	38	3	2							27	20	1	25	2	2	27	38	3	38	3	3
28	0	0	0	0	0							28	26	2	48	4	3	28	33	3	38	3	3
29	21	1	20	1	1							29	25	2	29	2	2	29	43	4	45	4	4
30	20	1	20	1	1							30	60	5	45	4	4	30	59	5	40	3	4
31	22	1	28	2	2							31	33	3	40	3	3	31	68	6	48	4	5
32	20	1	15	1	1							32	32	3	30	2	2	32	57	5	50	4	5
																		33	55	5	49	4	4
																		34	47	4	49	4	4
																		35	60	5	50	4	5
																		36	35	3	38	3	3
																		37	50	4	48	4	4

of the first judge show a deviation from the standards of 0 in the fourth grade, of —2 in the fifth grade, of —1 in the sixth grade, and of zero in the seventh grade.

Those of the second judge show deviations of —5, —2, 0, and +2, for the fourth, fifth, sixth, and seventh grades, respectively.

TABLE 7

DISTRIBUTION BY GRADES

Steps	0	1	2	3	4	5	6	7	8	9
Grade IV	1	13	17	0	1	0	0	0	0	0
Grade V		5	10	9	2	0	0	0	0	0
Grade VI			10	14	6	2	0	0	0	0
Grade VII				17	12	8	0	0	0	0

Table 9 shows a distribution of the papers by the individual judges.

Table 6 shows that the widest deviation in qualities assigned to any one paper by the individual judges is from

TABLE 8

MEDIANS OF GRADES

Grade	IV	V	VI	VII
Standard median	26	31	36	41
Grade median	23	30	36	44
First judge's median	26	29	35	41
Second judge's median	21	28	36	43

zero to 26; but it also shows that in many cases, there is no deviation whatever, or very slight deviation. This indicates that although a wide range of individual judgments is possible, the scale is an aid to the judgment in rating compositions.

It is true that, because of the impossibility of eliminating subjective reactions, one constantly feels a tendency to throw aside the scale, and to use personal judgment instead; but comparison of the medians in this test with the standard

medians will show that, although the scale is not a substitute for judgment, it can be used by them as a guide in rating the scales.

TABLE 9

DISTRIBUTION BY INDIVIDUAL JUDGES

1. *Grade IV*

Steps	0	1	2	3	4	5	6	7	8	9
By first judge	1	5	25	1	0	0	0	0	0	0
By second judge	2	15	13	1	1	0	0	0	0	0

2. *Grade V*

By first judge	0	6	10	9	1	0	0	0	0	0
By second judge	0	5	13	5	3	0	0	0	0	0

3. *Grade VI*

By first judge	0	3	9	13	1	6	0	0	0	0
By second judge	0	0	11	11	10	0	0	0	0	0

4. *Grade VII*

By first judge	0	0	2	20	5	8	1	1	0	0
By second judge	0	0	0	18	12	7	0	0	0	0

XII

THE USE OF THE BALLOU SCALE ON A SET OF COMPOSITIONS WRITTEN BY SEVENTH GRADE PUPILS

GRACE E. MANSON AND LOUISE W. LINTHICUM

Educational scales have developed out of actual school experience and in response to school needs. A study of the practice of teachers in marking discloses: (1) wide variability of standards from subject to subject and from school to school; (2) a need of more definite and concrete standards by which to measure school work. The purpose of the Ballou scale is to create an objective standard for measuring English compositions in order to make the judgments of English teachers more uniform. This objective standard shall serve as a basis for the exercise of subjective judgment.

The complete scale is composed of four separate scales: one for narration, a second for description, a third for exposition, and a fourth for argumentation. Each of these scales is composed of the type compositions. The subject of each is different. They are ranked approximately 95 per cent, 85 per cent, 75 per cent, 65 per cent, 55 per cent, 45 per cent. Under each composition is a series of remarks made by the compilers under these headings: "Merits," which tell the weak points: "Comparison," which justifies the position of the given composition in the scale. Each composition with rating is intended as an objective measure for any composition work of eighth grade pupils. The compilers believe it can be used to measure seventh and ninth grade work as well.

In order to use the scale:

1. Find to which style of discourse the composition belongs.

2. By comparison with the scale, roughly divide the compositions into six groups accrediting to them relative merits as measured by the six types of the scale.

3. Then grade them in the class to which they belong. For example, if there were five composition in A class, they might be graded 93, 91, 90, 89, 87, according to their individual merits and defects as measured by the scale, and as compared with each other.

The report which follows is an account of the use of this scale on a set of compositions written by seventh grade pupils in the Johns Hopkins Demonstration School.

The two students assigned to give the test selected six suitable descriptive topics, after which one was to be selected by the writers. In order that the class might be as little disturbed as possible, the grade teacher was asked to have the composition written. The class chose "A Fire Engine House." The time used was twenty-two minutes. Thirty-four compositions were written.

Each of the teachers making the test graded the papers independently by the scale. As a comparative study, a class of twenty-four English teachers in the Hopkins Summer School were asked to mark the papers by the ordinary percentage method.

1. To find the range of variations made by the class using the percentile method. (See Table 10.)

2. To find the average grade given the papers by each of the twenty-four readers. (See Table 11.)

3. To find the coefficients of correlation between the average ranking of the class and each investigator; also the coefficients of the two investigators. (See following paragraph 6.)

4. To find and compare the median grade of the class with the medians of each of the users of the scale. (See following paragraph 5.)

5. To check the scores made in this test with the scores made in the Port Townsend, Washington, test. (See following paragraph 4.)

TABLE 10

Scores of Judges

Reader	No. of Composition																															Readers Average
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	
1	80	90	70	80	75	95	65	90	70	75	80	90	75	75	85	85	65	60	30	60	95	80	85	60	70	70	55	80	80	75	50	65	85	70	71.61
2	60	70	90	70	40	95	65	85	75	60	80	95	60	60	70	70	50	70	20	95	60	60	75	80	70	75	75	75	70	75	25	50	65	75	68.5
3	75	60	68	77	70	73	80	80	78	78	55	75	75	60	68	80	65	67	5	70	90	55	76	40	50	73	45	65	79	85	75	42	60	78	68.33
4	80	85	80	60	25	90	60	60	60	50	60	92	80	83	70	65	60	60	15	70	90	65	80	65	83	75	60	75	75	80	70	50	60	75	67.29
5	50	80	85	60	60	80	50	40	90	70	50	95	50	50	70	50	80	40	10	70	30	30	60	70	75	80	20	70	30	20	20	50	40	40	54.85
6	60	80	80	79	65	72	48	76	80	35	45	84	74	75	79	79	5	55	10	48	85	80	60	70	79	82	65	78	61	75	20	75	80	70	69.08
7	65	65	75	56	48	70	68	65	68	70	50	75	60	68	68	65	75	50	10	50	68	68	60	66	65	70	65	70	60	50	20	55	70	50	60.00
8	75	92	94	80	70	95	60	84	60	88	78	94	80	82	88	80	75	75	25	74	80	80	74	80	78	94	70	95	85	84	80	70	80	80	78.50
9	70	80	90	75	60	80	60	70	70	60	72	90	88	85	75	80	75	60	20	60	90	70	80	80	72	75	60	80	85	72	50	50	65	80	71.40
10	60	80	65	80	75	85	65	70	75	75	65	75	70	65	70	70	65	50	5	70	78	60	65	60	75	75	70	70	60	75	25	50	65	75	65.5
11	80	70	70	70	65	85	70	75	76	75	65	78	50	60	70	80	55	60	10	70	65	60	80	75	76	75	72	75	60	88	70	55	65	75	68.5
12	85	85	95	70	73	92	85	80	80	74	83	95	83	90	70	90	72	70	20	65	90	80	74	75	72	75	58	72	80	90	88	70	75	70	77.14
13	60	85	90	85	40	80	60	70	60	40	70	90	75	75	65	70	60	65	25	50	80	55	75	60	80	60	70	85	85	80	30	75	50	60	62.
14	75	65	85	55	40	75	80	50	40	40	70	75	60	60	75	60	55	50	0	60	70	50	70	50	80	80	70	65	75	60	50	60	60	75	61.
15	65	80	80	75	56	81	50	60	70	75	76	85	60	65	80	75	75	55	10	60	75	75	70	60	56	75	56	80	78	60	50	55	60	75	66.4
16	60	78	90	60	50	60	65	78	35	78	50	88	60	80	75	79	70	50	0	25	73	50	50	60	85	75	58	80	73	75	15	30	65	50	60.9
17	40	80	90	85	60	85	50	85	75	70	70	95	80	75	85	85	75	60	5	80	85	65	80	70	75	70	75	70	85	80	60	65	70	80	72.3
18	60	75	90	65	40	75	55	90	55	60	65	85	85	80	75	70	60	65	0	40	75	60	75	50	80	85	55	90	80	60	65	70	70	65	66.07
19	75	60	70	75	70	60	60	80	55	65	65	85	75	65	60	75	55	60	10	65	65	60	75	60	65	85	60	60	85	70	55	60	60	88	65.
20	72	75	85	70	65	88	65	88	70	65	60	85	80	65	72	65	72	60	10	65	80	70	75	68	68	80	62	85	73	88	50	60	65	65	69.5
21	60	85	84	75	70	90	55	65	60	80	60	85	80	75	75	55	75	40	20	40	85	80	45	70	70	82	40	80	60	90	40	65	80	60	66.6
22	50	80	90	60	70	90	85	80	85	20	80	95	85	90	80	83	80	75	20	70	80	85	60	83	75	70	70	93	70	88	50	50	65	70	72.85
23	60	80	83	80	75	80	80	65	78	75	72	80	65	75	85	82	50	70	0	50	82	78	83	74	78	75	68	75	80	80	20	55	75	85	70.00
24	60	80	85	65	65	90	60	85	70	75	70	80	70	75	88	55	60	50	0	40	85	65	50	60	88	65	75	65	65	90	51	60	65	75	67.10

TABLE 11

Range of Scores

	No. of Composition															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	85	92	95	85	75	95	85	90	90	88	83	95	88	90	88	90
	85	90	94	85	75	95	85	90	85	90	80	95	85	90	88	85
	80	85	90	85	75	95	80	88	80	78	80	95	85	85	85	85
	80	85	90	80	70	92	80	85	80	78	80	95	83	83	85	83
	75	85	90	80	70	90	80	85	78	75	78	95	80	82	85	82
	75	85	90	80	70	90	70	85	78	75	76	94	80	80	80	80
	75	80	90	79	70	90	68	84	76	75	72	92	80	80	80	80
	75	80	90	79	70	90	65	80	75	75	72	90	80	75	79	80
	72	80	90	75	70	88	65	80	75	75	70	90	80	75	75	80
	70	80	85	75	65	85	65	80	75	75	70	90	75	75	75	79
	65	80	85	75	65	85	65	80	70	74	70	88	75	75	75	79
	65	80	85	75	65	85	65	78	70	30	70	85	75	75	75	75
	60	80	85	70	65	81	60	76	70	70	65	85	75	75	75	75
	60	80	84	70	60	80	60	75	70	70	65	85	74	75	72	70
	60	80	83	70	60	80	60	70	70	65	65	85	70	68	70	70
	60	78	80	70	60	80	60	70	68	65	65	85	70	65	70	70
	60	75	80	65	56	80	60	70	62	60	60	84	65	65	70	70
	60	75	80	65	50	75	60	65	60	60	60	80	60	65	70	65
	60	70	75	60	48	75	55	65	60	60	60	80	60	65	70	65
	60	70	70	60	40	73	55	65	55	50	55	78	60	60	70	65
	60	65	70	60	40	72	50	60	55	40	50	75	60	60	68	60
	50	65	70	60	40	70	50	60	50	40	50	75	60	60	68	55
	50	60	65	56	40	60	50	50	40	35	50	75	50	60	65	55
	40	60	60	55	25	60	48	40	35	20	45	75	50	50	60	50
Class ave.....	65.9	79.5	82.6	71.3	59.3	81.9	64.2	73.7	67.7	64.7	63.2	86.2	71.6	72.2	74.9	72.8
Range of var....	45	32	30	30	50	35	37	50	55	68	38	20	38	40	28	40

No. of Composition																
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
75	30	80	95	85	85	83	88	94	75	95	85	90	88	75	85	88
75	25	75	90	80	83	80	85	85	75	93	85	90	80	75	80	85
70	25	74	90	80	80	80	83	85	75	90	85	90	75	70	80	80
70	20	70	90	80	80	80	80	82	72	85	85	88	70	70	80	80
70	20	70	90	80	80	75	80	82	70	85	85	88	70	70	75	80
67	20	70	85	80	80	75	80	80	70	80	80	88	65	65	75	78
65	20	70	85	78	76	74	79	80	70	80	86	85	60	65	70	75
65	20	70	85	75	75	70	78	80	70	80	80	84	55	65	70	75
60	15	65	85	70	75	70	78	75	70	80	80	80	51	60	70	75
60	10	65	82	70	75	70	76	75	68	80	79	80	50	60	65	75
60	10	65	80	68	75	70	75	75	65	78	78	80	50	60	65	75
60	10	60	80	65	75	68	75	75	65	75	75	80	50	60	65	75
60	10	60	80	65	74	66	75	75	62	75	75	75	50	55	65	75
60	10	60	80	65	74	65	75	75	60	75	73	75	50	55	65	70
53	10	60	78	60	70	60	72	75	60	75	73	75	50	55	65	70
50	10	60	75	60	70	60	72	75	60	72	70	75	40	55	65	70
50	5	50	75	60	65	60	70	75	58	70	70	75	30	50	65	70
50	5	50	73	60	60	60	70	73	58	70	65	72	25	50	60	65
50	5	50	70	60	60	60	70	70	56	70	61	70	25	50	60	65
50	0	48	68	55	60	60	68	70	55	70	60	60	20	50	60	60
50	0	40	66	55	60	50	65	70	55	65	60	60	20	50	60	60
50	0	40	65	50	50	50	65	70	45	65	60	60	20	50	60	50
40	0	40	60	50	50	50	56	65	40	65	60	50	20	42	50	50
40	0	25	30	30	45	40	50	60	20	60	30	20	15	30	40	40
58.4	11.6	59.0	77.3	65.8	70.2	66.0	73.5	75.8	61.4	76.3	72.2	74.5	47.0	57.7	68.4	70.2
35	30	55	65	55	40	43	38	34	55	35	55	70	73	45	45	48

Class ave.
Range of var

1. The range of variation for the twenty-four readers was found to be very wide. The greatest was 73 points in No. 31; the highest mark given it being 88 per cent; the lowest, 15 per cent. The least variation was twenty points in No. 12. The problem here seems to be: What standards, if any, had these twenty-four teachers in mind? Does this show need of better standardization of the judgments of teachers?

2. There was no agreement as to the best composition. The class, as an average, considered No. 12 best, with a grade of 86; one writer took No. 26 and valued it at 86; the other chose No. 3, and graded it at 80 per cent. The two testers and the twenty-four readers chose No. 19 as the poorest.

3. The averages given by the twenty-four readers, not marking by the scale, were very generally higher. These marks made an average score of 76.2; the average of one tester scored 61.23; the average of the other tester scored 49.2.

4. Starch, in his Educational Measurements, says that in the Port Townsend, Washington, test, the following scores were made:

Grade	5	6	7	8	10	11	12
Score	46	46	53	58	63	70	73

Comparing these scores with those obtained in the present test, it would seem that the twenty-four readers rated them as eleventh grade work; one teacher rated them as ninth grade work, and the other rated them as sixth grade work.

5. The median of the class was found to be 70.5; of one investigator 59; of the other 60.

6. The coefficient of correlation between the class and each investigator was found to be higher than the correlation between the investigators. The correlation between Investigator 1 and class was .41. The correlation between Investigator 2 and class was .56. The correlation between the two investigators was .26. The correlation was found by Spearman formula.

The poor correlation between the two investigators is probably due to three main causes:

1. The fact that this was the first time either tester had used the scale.

2. The fact that one was more experienced, and had a correspondingly keener judgment in evaluating compositions.

3. Certain defects in the scale itself. Among the obvious defects are:

(*a*) Lack of directions for giving the test. In this case a serious complication arose. In the conference over the respective marks, it was found that No. 31 was not on the subject given. One graded it by the scale according to its value as a composition; the other gave it zero because it was off subject. It was decided that the composition would have to be graded by the scale irrespective of whether or not it was on the subject.

(*b*) The scale does not tell what merits were considered, or whether or not all defects were considered of equal value.

The two students making this test agree (1) that a scale is of the highest value; (2) they think that the Ballou scale has obvious merits, and just as obvious limitations. It is good in that it limits its range to measuring work of one grade, and in that it has a scale for all four forms of this course. (3) They think that it is perfectly clear that in this case almost as much subjective element must have entered into their markings by the scale as in the class markings by percentile methods. They believe that the continued use of this scale would fix a more definite standard in their own minds, and with repeated use, their variation would be removed. (4) They feel, too, that this removal of the personal element altogether would not be a good thing. In the case quoted above, the child gained a grading even when intentionally or unintentionally passing off a substitute for the real thing. We think that the child was the loser morally and educationally.

4. It is harder to mark by the scale. It takes more time until the scale becomes absolutely a fixture in your mind.

5. The "A" class of composition seems too high.

6. This scale would not be a fair basis of comparison for two schools or two systems unless the test were given under conditions as nearly like the conditions under which the original compositions were written as possible.

The writers make the following suggestions:

1. Have a set of directions for giving the test.

2. Have the compilers explain what they are looking for; then grade a system of papers as a guide.

3. Have a series of compositions on the same subject. This would illustrate the degree of marking better and be less confusing.

4. To compile a scale for the four years of high school from material gathered from schools all over the United States rather than from one community.